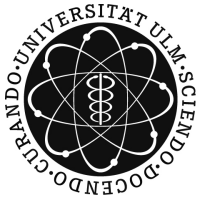


ulm university universität
uulm

Angewandte Stochastik

Vorlesungsskript

Prof. Dr. Evgeny Spodarev

Ulm

Sommersemester 2024

Vorwort

Das vorliegende Skript der Vorlesung Angewandte Stochastik gibt eine Einführung in die Problemstellungen der Wahrscheinlichkeitstheorie und der Statistik für Studierende der nicht mathematischen (jedoch mathematisch arbeitenden) Studiengänge wie Elektrotechnik, Informatik, Physik, usw. Es entstand in den Jahren 2005–2024, in denen ich diesen Vorlesungskurs an der Universität Ulm mehrmals gehalten habe.

Ich bedanke mich bei Herrn Tobias Scheinert, Herrn Michael Wiedler und Herrn Ly Viet Hoang für die Umsetzung meiner Vorlesungsnotizen in LaTeX.

Ulm, den 24. Juli 2024
Evgeny Spodarev

Inhaltsverzeichnis

Inhaltsverzeichnis i

1 Einführung 1
- 1.1 Über den Begriff “Stochastik” . . . 1
- 1.2 Geschichtliche Entwicklung der Stochastik . . . 2
- 1.3 Typische Problemstellungen der Stochastik . . . 5

2 Wahrscheinlichkeiten 6
- 2.1 Ereignisse . . . 7
- 2.2 Wahrscheinlichkeitsräume . . . 10
- 2.3 Beispiele . . . 12
  - 2.3.1 Klassische Definition der Wahrscheinlichkeiten . . . 13
  - 2.3.2 Geometrische Wahrscheinlichkeiten . . . 18
  - 2.3.3 Bedingte Wahrscheinlichkeiten . . . 19

3 Zufallsvariablen 26
- 3.1 Definition und Beispiele . . . 26
- 3.2 Verteilungsfunktion . . . 27
- 3.3 Grundlegende Klassen von Verteilungen . . . 30
  - 3.3.1 Diskrete Verteilungen . . . 30
  - 3.3.2 Absolut stetige Verteilungen . . . 34
  - 3.3.3 Mischungen von Verteilungen . . . 40
- 3.4 Verteilungen von Zufallsvektoren . . . 41
- 3.5 Stochastische Unabhängigkeit . . . 46
  - 3.5.1 Unabhängige Zufallsvariablen . . . 46
- 3.6 Funktionen von Zufallsvektoren . . . 48

4 Momente von Zufallsvariablen 53
- 4.1 Erwartungswert . . . 54
- 4.2 Varianz . . . 57
- 4.3 Kovarianz und Korrelationskoeffizient . . . 60
- 4.4 Höhere und gemischte Momente . . . 61
- 4.5 Entropie . . . 63
- 4.6 Ungleichungen . . . 65

5 Grenzwertsätze **67**
5.1 Gesetze der großen Zahlen 67
5.1.1 Schwaches und Starkes Gesetz der großen Zahlen . . . 68
5.1.2 Anwendung der Gesetze der großen Zahlen 69
5.2 Zentraler Grenzwertsatz 71
5.2.1 Klassischer zentraler Grenzwertsatz 71
5.2.2 Konvergenzgeschwindigkeit im zentralen Grenzwertsatz 73
5.2.3 Grenzwertsatz von Lindeberg 74

6 Monte–Carlo–Simulation von Zufallsvariablen **75**
6.1 Pseudozufallszahlen 75
6.2 Inversionsmethode 78
6.3 Akzeptanz– und Verwerfungsmethode 79
6.4 Simulation der Normalverteilung 81
6.4.1 Akzeptanz– und Verwerfungsmethode für $N(0,1)$ . . . 82
6.4.2 Box–Muller-Transformation 83
6.5 Simulation von diskret verteilten Zufallsvariablen 84
6.6 Markov-Ketten 87
6.6.1 Modellbeschreibung und Beispiele 88
6.6.2 Rekursive Darstellung und n-Schritt Übergangswahrscheinlichkeiten 91
6.6.3 Ergodizität von Markov-Ketten 94
6.6.4 Stationäre Anfangsverteilung und Reversibilität 100
6.7 Markov-Chain-Monte-Carlo-Simulation 103
6.7.1 Gibbs-Sampler 103
6.7.2 Metropolis-Hastings-Algorithmus 106

7 Beschreibende Statistik **110**
7.1 Typische Fragestellungen, Aufgaben und Ziele der Statistik . 110
7.2 Statistische Merkmale und ihre Typen 111
7.3 Statistische Daten und Stichproben 112
7.4 Stichprobenfunktionen 113
7.5 Verteilungen und ihre Darstellungen 114
7.5.1 Häufigkeiten und Diagramme 114
7.5.2 Empirische Verteilungsfunktion 116
7.6 Beschreibung von Verteilungen 117
7.6.1 Lagemaße 118
7.6.2 Streuungsmaße 122
7.6.3 Maße für Schiefe und Wölbung 124
7.7 Quantilplots (Quantil-Grafiken) 125
7.8 Dichteschätzung 129
7.9 Beschreibung und Exploration von bivariaten Datensätzen . . 131
7.9.1 Zusammenhangsmaße 131
7.9.2 Einfache lineare Regression 135

8 Punktschätzer **142**
8.1 Parametrisches Modell 142
8.2 Parametrische Familien von statistischen Prüfverteilungen . . 143
8.2.1 Gamma-Verteilung 143
8.2.2 Student-Verteilung (t-Verteilung) 146
8.2.3 Fisher-Snedecor-Verteilung (F-Verteilung) 146
8.3 Punktschätzer und ihre Grundeigenschaften 147
8.3.1 Eigenschaften von Punktschätzern 148
8.3.2 Schätzer des Erwartungswertes und empirische Momente 150
8.3.3 Schätzer der Varianz 150
8.3.4 Eigenschaften der Ordnungsstatistiken 153
8.3.5 Empirische Verteilungsfunktion 154

Literatur **157**

Kapitel 1

Einführung

1.1 Über den Begriff "Stochastik"

Wahrscheinlichkeitsrechnung ist eine Teildisziplin von Stochastik. Dabei kommt das Wort "Stochastik" aus dem Griechischen $\sigma\tau\omega\chi\alpha\sigma\tau\iota\kappa\eta$- "die Kunst des Vermutens" (von $\sigma\tau\omega\chi\omega\xi$ - "Vermutung, Ahnung, Ziel"). Dieser Begriff wurde von Jacob Bernoulli in seinem Buch "Ars conjectandi" geprägt (1713), in dem das erste Gesetz der großen Zahlen bewiesen wurde.

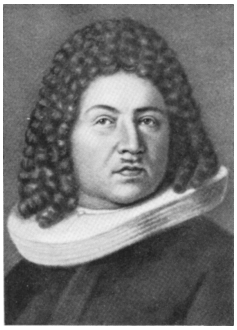


Abbildung 1.1: Jacob Bernoulli (1654-1705)

Stochastik beschäftigt sich mit den Ausprägungen und quantitativen Merkmalen von Zufall. Aber was ist Zufall? Gibt es Zufälle überhaupt? Das ist eine philosophische Frage, auf die jeder seine eigene Antwort suchen muss. Für die moderne Mathematik ist der Zufall eher eine Arbeitshypothese, die viele Vorgänge in der Natur und in der Technik ausreichend gut zu beschreiben scheint. Insbesondere kann der Zufall als eine Zusammenwirkung mehrerer Ursachen aufgefasst werden, die sich dem menschlichen Verstand entziehen (z.B. Brownsche Bewegung). Andererseits gibt es Studienbereiche (wie z.B. in der Quantenmechanik), in denen der Zufall eine essentielle Rolle zu spielen scheint (die Unbestimmtheitsrelation von Heisenberg). Wir werden die Existenz des Zufalls als eine wirkungsvolle Hypothese annehmen, die für viele Bereiche des Lebens zufriedenstellende Antworten liefert.

Stochastik kann man in folgende Gebiete unterteilen:

- Wahrscheinlichkeitsrechnung oder Wahrscheinlichkeitstheorie (Grundlagen)
- Statistik (Umgang mit den Daten)
- Stochastische Prozesse (Theorie zufälliger Zeitreihen und Felder)

- Simulation

Diese Vorlesung ist nur dem ersten Teil gewidmet.

1.2 Geschichtliche Entwicklung der Stochastik

1. *Vorgeschichte:*
   Die Ursprünge der Wahrscheinlichkeitstheorie liegen im Dunklen der alten Zeiten. Ihre Entwicklung ist in der ersten Phase den Glücksspielen zu verdanken. Die ersten Würfelspiele konnte man in Altägypten, I. Dynastie (ca. 3500 v. Chr.) nachweisen. Auch später im klassischen Griechenland und im römischen Reich waren solche Spiele Mode (Kaiser August (63 v. Chr. -14 n. Chr.) und Claudius (10 v.Chr. - 54 n. Chr.).
   Gleichzeitig gab es erste wahrscheinlichkeitstheoretische Überlegungen in der Versicherung und im Handel. Die älteste uns bekannte Form der Versicherungsverträge stammt aus dem Babylon (ca. 4-3 T. J. v.Chr., Verträge über die Seetransporte von Gütern). Die ersten Sterbetafeln in der Lebensversicherung stammen von dem römischen Juristen Ulpian (220 v.Chr.). Die erste genau datierte Lebensversicherungspolice stammt aus dem Jahre 1347, Genua.

   Abbildung 1.2: Gerolamo Cardano (1501-1576)

   Der erste Wissenschaftler, der sich mit diesen Aufgabenstellungen aus der Sicht der Mathematik befasst hat, war G. Cardano, der Erfinder der Cardan–Welle. In seinem Buch “Liber de ludo alea” sind zum ersten Mal Kombinationen von Ereignissen beim Würfeln beschrieben, die vorteilhaft für den Spieler sind. Er hat auch als erster die

   $$\frac{\text{Anzahl vorteilhafter Ereignisse}}{\text{Anzahl aller Ereignisse}}$$

   als Maß für Wahrscheinlichkeit entdeckt.

2. *Klassische Wahrscheinlichkeiten (XVII-XVIII Jh.):*
   Diese Entwicklungsperiode beginnt mit dem Briefwechsel zwischen Blaise Pascal und Pierre de Fermat. Sie diskutierten Probleme, die von Chevalier de Méré (Antoine Gombaud (1607-1684)) gestellt wurden. Anbei ist eines seiner Probleme:
   Was ist wahrscheinlicher: mindestens eine 6 bei 4 Würfen eines Würfels oder mindestens ein Paar (6,6) bei 24 Würfen von 2 Würfeln zu bekommen?

A = {mind. eine 6 bei 4 Würfen eines Würfels}
B = {mind. (6,6) bei 24 Würfen von zwei Würfeln}

Die Antwort:

P(mind. eine 6 in 4 Würfen)
$= P(A) = 1 - P(\overline{A}) = 1 - P$(keine 6)
$= 1 - \left(\frac{5}{6}\right)^4 = 0,516 > 0,491 = 1 - \left(\frac{35}{36}\right)^{24}$
$= 1 - P(\overline{B}) = P(B) = P$(mind. 1 (6,6) in 24 Würfen von 2 Würfeln)

Abbildung 1.3: Blaise Pascal (1623-1662), Pierre de Fermat (1601-1665) und Christian Huygens (1629-1695)

Weitere Entwicklung:

1657	Christian Huygens "De Ratiociniis in Ludo Alea" (Operationen mit Wahrscheinlichkeiten)
1713	Jacob Bernoulli "Ars Conjectandi" (Wahrscheinlichkeit eines Ereignisses und Häufigkeit seines Eintretens)

3. *Entwicklung analytischer Methoden (XVIII-XIX Jh.)* von Abraham de Moivre, Thomas Bayes (1702-1761), Pierre Simon de Laplace, Carl Friedrich Gauß, Simeon Denis Poisson (vgl. Abb. 1.4).
Entwicklung der Theorie bezüglich Beobachtungsfehlern und der Theorie des Schießens (Artilleriefeuer). Erste nicht–klassische Verteilungen wie Binomial– und Normalverteilung ($f(x) = \frac{1}{\sqrt[2]{2\pi}} e^{-}(\frac{x^2}{2}), x \in \mathbb{R}$), Poisson–Verteilung, zentraler Grenzwertsatz von De Moivre.

St.–Petersburger Schule von Wahrscheinlichkeiten:
(P.L. Tschebyschew, A.A. Markow, A.M. Ljapunow)

– Einführung von Zufallsvariablen, Erwartungswerten, Wahrscheinlichkeitsfunktionen, Markow–Ketten, abhängigen Zufallsvariablen.

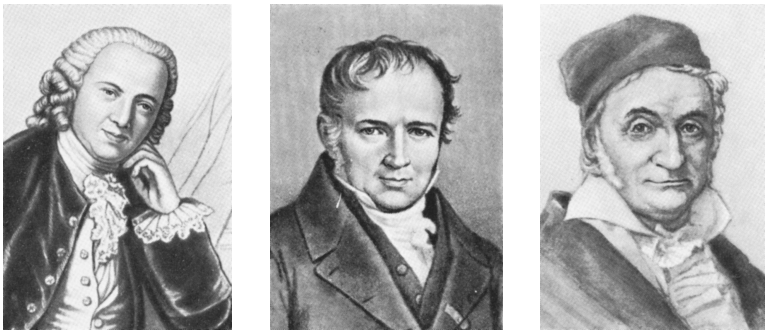


Abbildung 1.4: Abraham de Moivre (1667-1754), Pierre Simon de Laplace (1749-1827) und Karl Friedrich Gauß (1777-1855)

Abbildung 1.5: Simeon Denis Poisson (1781-1840), P. L. Tschebyschew (1821-1894) und A. A. Markow (1856-1922)

4. *Moderne Wahrscheinlichkeitstheorie (XX Jh.) David Hilbert, 8.8.1900, II. Mathematischer Kongress in Paris, Problem Nr. 6:*
   Axiomatisierung von physikalischen Disziplinen, wie z.B. Wahrscheinlichkeitstheorie.
   R. v. Mises: frequentistischer Zugang: $P(A) = \lim_{n\to\infty} \frac{\#A \text{ in } n \text{ } Versionen}{n}$
   Antwort darauf: A.N. Kolmogorow führt Axiome der Wahrscheinlichkeitstheorie basierend auf der Maß– und Integrationstheorie von Borel und Lebesgue (1933) ein.

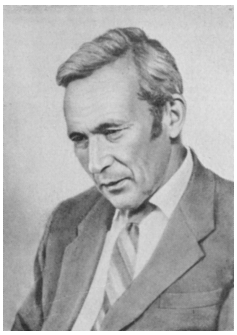
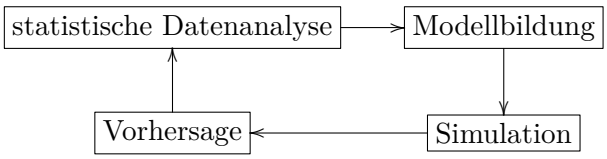


Abbildung 1.6: A. M. Ljapunow (1857-1918) und A. H. Kolmogorow (1903-1987)

1.3 Typische Problemstellungen der Stochastik



1. Modellierung von Zufallsexperimenten, d.h. deren adäquate theoretische Beschreibung.

2. Bestimmung von

   - Wahrscheinlichkeiten von Ereignissen
   - Mittelwerten und Varianzen von Zufallsvariablen
   - Verteilungsgesetzen von Zufallsvariablen

3. Näherungsformel und Lösungen mit Hilfe von Grenzwertsätzen

4. Schätzung von Modellparametern in der Statistik, Prüfung statistischer Hypothesen

Kapitel 2

Wahrscheinlichkeiten

Wahrscheinlichkeitstheorie befasst sich mit (im Prinzip unendlich oft) wiederholbaren Experimenten, in Folge derer ein Ereignis auftreten kann (oder nicht). Solche Ereignisse werden "zufällige Ereignisse" genannt. Sei A ein solches Ereignis. Wenn $n(A)$ die Häufigkeit des Auftretens von A in n Experimenten ist, so hat man bemerkt, dass $\frac{n(A)}{n} \to c$ für große n ($n \to \infty$). Diese Konstante c nennt man "Wahrscheinlichkeit von A" und bezeichnet sie mit $P(A)$.
Beispiel: *n–maliger Münzwurf:* (siehe Abbildung 2.1) faire Münze, d.h. $n(A) \approx n(\bar{A}), \qquad A = \{\text{Kopf}\}, \quad \bar{A} = \{\text{Zahl}\}$

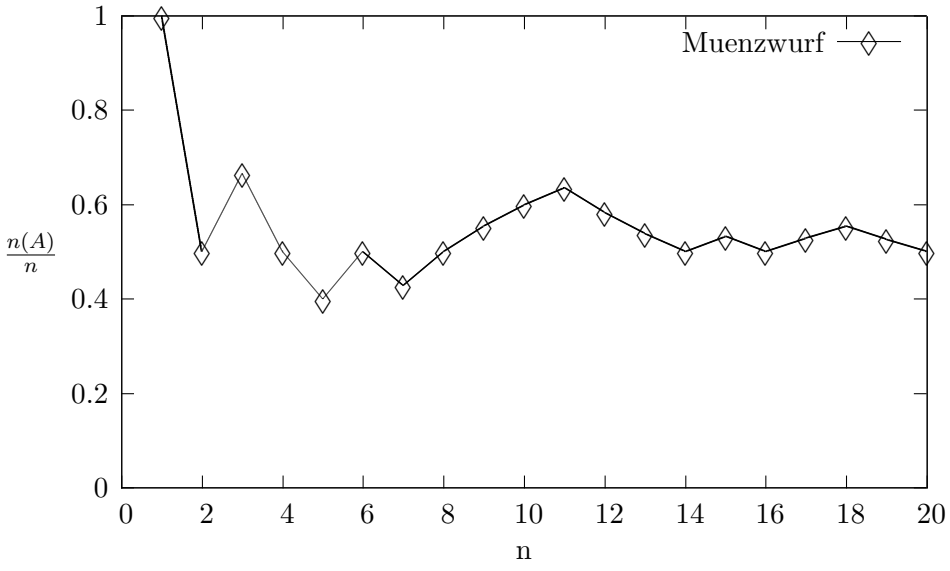


Abbildung 2.1: Relative Häufigkeit $\frac{n(A)}{n}$ des Ereignisses "Kopf" beim n–maligen Münzwurf

Man kann leicht feststellen, dass $\frac{n(A)}{n} \approx \frac{1}{2}$ für große n. $\Longrightarrow P(A) = \frac{1}{2}$.
Um dies zu verifizieren, hat Buffon in XVIII Jh. 4040 mal eine faire Münze geworfen, davon war 2048 mal Kopf, so dass $\frac{n(A)}{n} = 0,508$.
Pierson hat es 24000 mal gemacht: es ergab $n(A) = 12012$ und somit $\frac{n(A)}{n} \approx 0.5005$.

In den Definitionen, die wir bald geben werden, soll diese empirische Begriffsbildung $P(A) = \lim_{n\to\infty} \frac{n(A)}{n}$ ihren Ausdruck finden. Zunächst definieren wir, für welche Ereignisse A die Wahrscheinlichkeit $P(A)$ überhaupt eingeführt werden kann.

offene Fragen:

1. Was ist $P(A)$?
2. Für welche A ist $P(A)$ definiert?

2.1 Ereignisse

Sei E ein Grundraum und $\Omega \subseteq E$ sei die Menge von Elementarereignissen ω (Grundmenge).
Ω kann als Menge der möglichen Versuchsergebnisse interpretiert werden. Man nennt Ω manchmal auch *Grundgesamtheit* oder *Stichprobenraum*.

Definition 2.1.1 Eine Teilmenge A von Ω ($A \subset \Omega$) wird *Ereignis* genannt. Dabei ist $\{\omega\} \subset \Omega$ ein *Elementarereignis*, das das Versuchsergebnis ω darstellt. Falls bei einem Versuch das Ergebnis $\omega \in A$ erzielt wurde, so sagen wir, dass A eintritt.

Beispiel 2.1.1

1. *Einmaliges Würfeln:* $\Omega = \{1, 2, 3, 4, 5, 6\}, \quad E = \mathbb{N}$
2. n–maliger Münzwurf: $\Omega = \{(\omega_1, \ldots, \omega_n) : \omega_i \in \{0, 1\}\}$
   $E = \mathbb{N}^n$, $\omega_i = \begin{cases} 1, & \text{falls ein “Kopf” im i–ten Wurf} \\ 0, & \text{sonst} \end{cases}$

Weiter werden wir E nicht mehr spezifizieren.

Tabelle 2.1: Wahrscheinlichkeitstheoretische Bedeutung von Mengenoperationen

$A = \emptyset$	unmögliches Ereignis
$A = \Omega$	wahres Ereignis
$A \subset B$	aus dem Eintreten von A folgt auch, dass B eintritt.
$A \cap B = \emptyset$	(disjunkte, *unvereinbare* Ereignisse): A und B können nicht gleichzeitig eintreten.
$A = B \cup C$	Mindestens eines der Ereignisse B und C tritt ein.
$A = \cup_{i=1}^{n} A_i$	Ereignis $A =$ "Es tritt mindestens ein Ereignis A_i ein"
$A = B \cap C$	Ereignis $A =$ "Es treten B und C gleichzeitig ein."
$A = \cap_{i=1}^{n} A_i$	Ereignis $A =$ "Es treten alle Ereignisse $A_1, \ldots, A_n$ ein"
$\bar{A} = A^c$	Das Ereignis A tritt nicht ein.
$A = B \backslash C$	Ereignis A tritt genau dann ein, wenn B eintritt, aber nicht C
$A = B \triangle C$	Ereignis A tritt genau dann ein, wenn B oder C eintreten (*nicht gleichzeitig!*)

Anmerkung: $A = B \triangle C = (B \backslash C) \cup (C \backslash B)$ (symmetrische Differenz)

Definition 2.1.2 Ereignisse $A_1, A_2, A_3, \ldots$ heißen *paarweise disjunkt* oder *unvereinbar*, wenn $A_i \cap A_j = \emptyset \quad \forall i \neq j$

Beispiel 2.1.2 *Zweimaliges Würfeln*: $\Omega = \{(\omega_1, \omega_2) : \omega_i \in \{1, \ldots, 6\}, i = 1, 2\}$,
$A =$ "die Summe der Augenzahlen ist 6" $= \{(1,5), (2,4), (3,3), (4,2), (5,1)\}$.
Die Ereignisse $B =$ "Die Summe der Augenzahlen ist ungerade" und
$C = \{(3,5)\}$ sind unvereinbar.

Oft sind nicht alle Teilmengen von Ω als Ereignisse sinnvoll. Deswegen beschränkt man sich auf ein Teilsystem von Ereignissen mit bestimmten Eigenschaften; und zwar soll dieses Teilsystem abgeschlossen bezüglich Mengenoperationen sein.

Definition 2.1.3 Eine nicht leere Familie $\mathcal{F}$ von Ereignissen aus Ω heißt *Algebra*, falls

1. $A \in \mathcal{F} \Longrightarrow \bar{A} \in \mathcal{F}$
2. $A, B \in \mathcal{F} \Longrightarrow A \cup B \in \mathcal{F}$

Beispiel 2.1.3 1. Die *Potenzmenge* $\mathcal{P}(\Omega)$ = (die Menge aller Teilmengen von Ω) ist eine Algebra.

2. Im Beispiel 2.1.2 ist $\mathcal{F} = \{\emptyset, A, \bar{A}, \Omega\}$ eine Algebra. Dagegen ist $\mathcal{F} = \{\emptyset, A, B, C, \Omega\}$ keine Algebra: z.B. $A \cup C \notin \mathcal{F}$.

Lemma 2.1.1 (*Eigenschaften einer Algebra:*) Sei $\mathcal{F}$ eine Algebra von Ereignissen aus Ω. Es gelten folgende Eigenschaften:

1. $\emptyset, \Omega \in \mathcal{F}$
2. $A, B \in \mathcal{F} \Longrightarrow A \backslash B \in \mathcal{F}$
3. $A_1, \ldots, A_n \in \mathcal{F} \Longrightarrow \bigcup_{i=1}^{n} A_i \in \mathcal{F}$ und $\bigcap_{i=1}^{n} A_i \in \mathcal{F}$

Beweis

1. $\mathcal{F} \neq \emptyset \Longrightarrow \exists A \in \mathcal{F} \Longrightarrow \bar{A} \in \mathcal{F}$ nach Definition $\Longrightarrow A \cup \bar{A} = \Omega \in \mathcal{F}$; $\emptyset = \bar{\Omega} \in \mathcal{F}$.
2. $A, B \in \mathcal{F}, \quad A \backslash B = A \cap \bar{B} = \overline{(\bar{A} \cup B)} \in \mathcal{F}$.
3. *Induktiver Beweis:*
   $n = 2$: $A, B \in \mathcal{F} \Longrightarrow A \cap B = \overline{(\bar{A} \cup \bar{B})} \in \mathcal{F}$
   $n = k \mapsto n = k + 1$: $\quad \bigcap_{i=1}^{k+1} A_i = (\bigcap_{i=1}^{k} A_i) \cap A_{k+1} \in \mathcal{F}$.

□

Für die Entwicklung einer gehaltvollen Theorie sind aber Algebren noch zu allgemein. Manchmal ist es auch notwendig, unendliche Vereinigungen $\bigcup_{i=1}^{\infty} A_i$ oder unendliche Schnitte $\bigcap_{i=1}^{\infty} A_i$ zu betrachten, um z.B. Grenzwerte von Folgen von Ereignissen definieren zu können. Dazu führt man Ereignissysteme ein, die *σ–Algebren* genannt werden:

Definition 2.1.4

1. Eine Algebra $\mathcal{F}$ heißt *σ–Algebra*, falls aus $A_1, A_2, \ldots, \in \mathcal{F}$ folgt, dass $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.
2. Das Paar $(\Omega, \mathcal{F})$ heißt *Messraum*, falls $\mathcal{F}$ eine σ–Algebra der Teilmengen von Ω ist.

Beispiel 2.1.4 1. $\mathcal{F} = \mathcal{P}(\Omega)$ ist eine σ–Algebra.

2. *Beispiel einer Algebra $\mathcal{F}$, die keine σ–Algebra ist:*
   Sei $\mathcal{F}$ die Klasse von Teilmengen aus $\Omega = \mathbb{R}$, die aus endlichen Vereinigungen von disjunkten Intervallen der Form $(-\infty, a], (b, c]$ und (d, ∞)

$a, b, c, d \in \mathbb{R}$, besteht. Offensichtlich ist $\mathcal{F}$ eine Algebra. Dennoch ist $\mathcal{F}$ keine σ–Algebra, denn $[b, c] = \cap_{n=1}^{\infty} \underbrace{(b - \frac{1}{n}, c]}_{\in \mathcal{F}} \notin \mathcal{F}$.

2.2 Wahrscheinlichkeitsräume

Auf einem Messraum $(\Omega, \mathcal{F})$ wird ein *Wahrscheinlichkeitsmaß* durch folgende *Axiome von Kolmogorow* eingeführt:

Definition 2.2.1 1. Die Mengenfunktion $P : \mathcal{F} \to [0, 1]$ heißt *Wahrscheinlichkeitsmaß* auf $\mathcal{F}$, falls

   (a) $P(\Omega) = 1$ *(Normiertheit)*

   (b) $\{A_n\}_{n=1}^{\infty} \subset \mathcal{F}$, wobei A_n paarweise disjunkt $\Longrightarrow P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$ *(σ–Additivität)*

2. Das Tripel $(\Omega, \mathcal{F}, P)$ heißt *Wahrscheinlichkeitsraum*.

3. $\forall A \in \mathcal{F}$ heißt $P(A)$ *Wahrscheinlichkeit* des Ereignisses A.

Bemerkung 2.2.1 1. Nachfolgend werden nur solche $A \subset \Omega$ *Ereignisse* genannt, die zu der ausgewählten σ–Algebra $\mathcal{F}$ von Ω gehören. Alle anderen Teilmengen $A \subset \Omega$ sind demnach *keine* Ereignisse.

2. $\mathcal{F}$ kann nicht immer als $\mathcal{P}(\Omega)$ gewählt werden. Falls Ω endlich oder abzählbar ist, ist dies jedoch möglich. Dann kann $P(A)$ auf $(\Omega, \mathcal{P}(\Omega))$ als $P(A) = \sum_{\omega \in A} P(\{\omega\})$ definiert werden (klassische Definition der Wahrscheinlichkeiten).
   Falls z.B. $\Omega = \mathbb{R}$ ist, dann kann $\mathcal{F}$ nicht mehr als $\mathcal{P}(\mathbb{R})$ gewählt werden, weil ansonsten $\mathcal{F}$ eine große Anzahl von *pathologischen* Ereignissen enthalten würde, für die z.B. der Begriff der Länge nicht definiert ist.

Definition 2.2.2 Sei $\mathcal{U}$ eine beliebige Klasse von Teilmengen aus Ω. Dann ist durch

$$\sigma(\mathcal{U}) = \bigcap_{\mathcal{U} \subset \mathcal{F},\, \mathcal{F} - \sigma - \text{Alg. von } \Omega} \mathcal{F}$$

eine σ–Algebra gegeben, die *minimale σ–Algebra, die U enthält*, genannt wird.

Übungsaufgabe 2.2.1 Zeigen Sie bitte, dass die in Definition 2.2.2 definierte Klasse $\sigma(U)$ tatsächlich eine σ–Algebra darstellt.

Definition 2.2.3 Sei $\Omega = \mathbb{R}^d$. Sei $\mathcal{U}$ = Klasse aller offenen Teilmengen von $\mathbb{R}^d$. Dann heißt $\sigma(\mathcal{U})$ die *Borel σ–Algebra* auf $\mathbb{R}^d$ und wird mit $\mathfrak{B}_{\mathbb{R}^d}$ bezeichnet. Elemente von $\mathfrak{B}_{\mathbb{R}^d}$ heißen *Borel–Mengen*. Diese Definition kann auch für einen beliebigen topologischen Raum Ω (nicht unbedingt $\mathbb{R}^d$) gegeben werden.

Übungsaufgabe 2.2.2 Zeigen Sie, dass $\mathcal{B}_{R^d}$ alle $\{x\}$, $x \in \mathbb{R}^d$, alle abgeschlossenen und insbesondere kompakten Teilmengen von $\mathbb{R}^d$ enthält, z.B. $[a,b]^d \in \mathcal{B}_{R^d}$, $a \leq b$.

Satz 2.2.1 Sei $(\Omega, \mathcal{F}, P)$ ein Wahrscheinlichkeitsraum und $A_1, \ldots, A_n, A$, $B \subset \mathcal{F}$. Dann gilt:

1. $P(\emptyset) = 0$.
2. $P(\bigcup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$, falls A_i paarweise disjunkt sind.
3. Falls $A \subset B$, dann ist $P(B \backslash A) = P(B) - P(A)$.

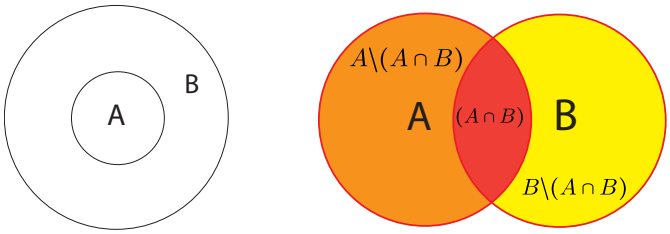


Abbildung 2.2: Illustration zu Satz 2.2.1, 3) und 4)

4. $P(A \cup B) = P(A) + P(B) - P(A \cap B), \quad \forall A, B \in \mathcal{F}$.
5. $P(\overline{B}) = 1 - P(B)$

Beweis

1. $\emptyset = \cup_{i=1}^{\infty} \emptyset \Longrightarrow P(\emptyset) = \sum_{i=1}^{\infty} P(\emptyset) \leq 1 \Longrightarrow P(\emptyset) = 0$
2. $P(\bigcup_{i=1}^n A_i) = P(\bigcup_{i=1}^n A_i \cup \bigcup_{k=n+1}^{\infty} \emptyset) = \sum_{i=1}^n P(A_i) + \sum_{i=n+1}^{\infty} 0 = \sum_{i=1}^n P(A_i)$
3. $P(B) = P(A \cup (B \backslash A)) = P(A) + P(B \backslash A) \Longrightarrow$ geht.
4. Benutze 2), 3) und $A \cup B = [A \backslash (A \cap B)] \cup [A \cap B] \cup [B \backslash (A \cap B)]$

$$\begin{aligned} P(A \cup B) &= P(A) - P(A \cap B) + P(A \cap B) + P(B) - P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

(vgl. Abb. 2.2).

5. $\overline{B} = \Omega \setminus B \Rightarrow P(\overline{B}) = \underbrace{P(\Omega)}_{=1} - P(B) = 1 - P(B)$.

□

Folgerung 2.2.1
Es gelten folgende Eigenschaften von P für $A_1, \ldots, A_n, A, B \in \mathcal{F}$:

1. $A \subseteq B \Longrightarrow P(A) \leq P(B)$ (Monotonie)
2. $P(\bigcup_{i=1}^n A_i) \leq \sum_{i=1}^n P(A_i), \quad \forall A_1, \ldots, A_n \in \mathcal{F}$ (Subadditivität)
3. *Siebformel:*

$$P(\cup_{i=1}^n A_i) = \sum_{i=1}^n (-1)^{i-1} \sum_{1 \leq k_1 < \cdots < k_i \leq n} P(A_{k_1} \cap \cdots \cap A_{k_i})$$

Beweis

1. (Folgt aus Satz 2.2.1) $P(B) = P(A) + \underbrace{P(B \backslash A)}_{\geq 0} \geq P(A)$
2. Induktion nach n: n=2: $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) \leq P(A_1) + P(A_2)$
3. Induktion nach n: Der Fall $n = 2$ folgt aus Satz 2.2.1, 4). Der Rest ist klar.

□

Übungsaufgabe 2.2.3 Führen Sie die Induktion bis zum Ende durch.

Die Eigenschaft $P(\bigcup_{n=1}^k A_n) \leq \sum_{n=1}^k P(A_n)$ heißt *Subadditivität des Wahrscheinlichkeitsmaßes* P. Diese Eigenschaft gilt jedoch auch für unendlich viele A_n, wie folgendes Korollar zeigt:

Folgerung 2.2.2 *σ–Subadditivität:* Sei $(\Omega, \mathcal{F}, P)$ ein Wahrscheinlichkeitsmaß und $\{A_n\}_{n \in \mathbb{N}}$ eine Folge von Ereignissen. Dann gilt $P(\bigcup_{n=1}^\infty A_n) \leq \sum_{n=1}^\infty P(A_n)$, wobei die rechte Seite nicht unbedingt endlich sei soll (dann ist die Aussage trivial).

Definition 2.2.4
1. Ereignisse A und B heißen *(stochastisch) unabhängig*, falls $P(A \cap B) = P(A) \cdot P(B)$.
2. Eine Folge von Ereignissen $\{A_n\}_{n \in \mathbb{N}}$ (diese Folge kann auch endlich viele Ereignisse enthalten!) heißt *(stochastisch) unabhängig in ihrer Gesamtheit*, falls $\forall n \, \forall i_1 < i_2 < \cdots < i_n$

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_n}) = \prod_{k=1}^n P(A_{i_k}).$$

2.3 Beispiele

In diesem Abschnitt betrachten wir die wichtigsten Beispiele für Wahrscheinlichkeitsräume $(\Omega, \mathcal{F}, P)$. Wir beginnen mit:

2.3.1 Klassische Definition der Wahrscheinlichkeiten

Hier wird ein Grundraum Ω mit $|\Omega| < \infty$ ($|A| = \#A$ – die Anzahl von Elementen in A) betrachtet. Dann kann $\mathcal{F}$ als $\mathcal{P}(\Omega)$ gewählt werden. Die klassische Definition von Wahrscheinlichkeiten geht von der Annahme aus, dass alle Elementarereignisse ω gleich wahrscheinlich sind:

Definition 2.3.1

1. Ein Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, P)$ mit $|\Omega| < \infty$ ist ein *endlicher Wahrscheinlichkeitsraum*.

2. Ein endlicher Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, P)$ mit $\mathcal{F} = \mathcal{P}(\Omega)$ und

$$\forall \omega \in \Omega \quad P(\{\omega\}) = \frac{1}{|\Omega|}$$

heißt *Laplacescher Wahrscheinlichkeitsraum*. Das eingeführte Maß heißt *klassisches* oder *laplacesches Wahrscheinlichkeitsmaß*.

Bemerkung 2.3.1 Für die klassische Definition der Wahrscheinlichkeit sind alle Elementarereignisse $\{\omega\}$ gleich wahrscheinlich:
$\forall \omega \in \Omega \quad P(\{\omega\}) = \frac{1}{|\Omega|}$. Nach der Additivität von Wahrscheinlichkeitsmaßen gilt:

$$P(A) = \frac{|A|}{|\Omega|} \quad \forall A \subset \Omega\,.$$

(Beweis: $P(A) = \sum_{\omega \in A} P(\{\omega\}) = \sum_{\omega \in A} \frac{1}{|\Omega|} = \frac{|A|}{|\Omega|}$). Dabei heißt

$$P(A) = \frac{\text{Anzahl günstiger Fälle}}{\text{Anzahl aller Fälle}}\,.$$

Beispiel 2.3.1

1. *Problem von Galilei:*
   Ein Landsknecht hat Galilei (manche sagen, es sei Huygens passiert) folgende Frage gestellt: Es werden 3 Würfel gleichzeitig geworfen. Was ist wahrscheinlicher: Die Summe der Augenzahlen ist 11 oder 12? Nach Beobachtung sollte 11 öfter vorkommen als 12. Doch ist es tatsächlich so?

   - Definieren wir den Wahrscheinlichkeitsraum $\Omega = \{\omega = (\omega_1, \omega_2, \omega_3) : \omega_i \in \{1, \ldots, 6\}\}$,
     $|\Omega| = 6^3 = 216 < \infty, \quad \mathcal{F} = \mathcal{P}(\Omega)$; sei

$$B := \{\text{Summe der Augenzahlen } 11\}$$

$$= \{\omega \in \Omega : \omega_1 + \omega_2 + \omega_3 = 11\}$$
$$C := \{\text{Summe der Augenzahlen 12}\}$$
$$= \{\omega \in \Omega : \omega_1 + \omega_2 + \omega_3 = 12\}.$$

- *Lösung des Landknechtes:* 11 und 12 können folgendermaßen in die Summe von 3 Summanden zerlegt werden:
  $11 = 1+5+5 = 1+4+6 = 2+3+6 = 2+4+5 = 3+3+5 = 3+4+4 \Longrightarrow |B| = 6 \Longrightarrow P(B) = \frac{6}{6^3} = \frac{1}{36}$
  $12 = 1+5+6 = 2+4+6 = 2+5+5 = 3+4+5 = 3+3+6 = 4+4+4 \Longrightarrow P(C) = \frac{6}{6^3} = \frac{1}{36}$.
  Dies entspricht jedoch nicht der Erfahrung.

Die Antwort von Galilei war, dass der Landsknecht mit nicht unterscheidbaren Würfeln gearbeitet hat, somit waren Kombinationen wie (1,5,5), (5,1,5) und (5,5,1) identisch und wurden nur einmal gezählt. In der Tat ist es anders: Jeder Würfel hat eine Nummer, ist also von den anderen Zwei zu unterscheiden. Daher gilt $|B| = 27$ und $|C| = 25$, was daran liegt, das (4,4,4) nur einmal gezählt wird. Also

$$P(B) = \frac{|B|}{|\Omega|} = \frac{27}{216} > P(C) = \frac{|C|}{|\Omega|} = \frac{25}{216}$$

2. *Geburtstagsproblem:* Es gibt n Studenten in einem Jahrgang an der Uni, die dieselbe Vorlesung Angewandte Stochastik I besuchen. Wie groß ist die Wahrscheinlichkeit, dass mindestens 2 Studenten den Geburtstag am selben Tag feiern? Sei $M = 365 =$ Die Anzahl der Tage im Jahr. Dann gilt

$$\Omega = \{\omega = (\omega_1, \ldots, \omega_n) : \omega_i \in \{1, \ldots, M\}\}, |\Omega| = M^n < \infty\,.$$

Sei

$$\begin{aligned} A_n &= \{\text{min. 2 Studenten haben am gleichen Tag Geb.}\} \subset \Omega\,, \\ A_n &= \{\omega \in \Omega : \exists i, j \in \{1, \ldots, n\}, i \neq j : \omega_i = \omega_j\}\,, \\ P(A_n) &= ? \end{aligned}$$

Ansatz: $P(A_n) = 1 - P(\bar{A}_n)$, wobei

$$\bar{A}_n = \{\omega \in \Omega : \omega_i \neq \omega_j \quad \forall i \neq j \text{ in } \omega = (\omega_1, \ldots, \omega_n)\}$$

$|\bar{A}_n| = M(M-1)(M-2)\ldots(M-n+1)$. Somit gilt

$$P(\bar{A}_n) = \frac{M(M-1)\ldots(M-n+1)}{M^n}$$

n	4	16	22	23	40	64
$P(A_n)$	0,016	0,284	0,476	0,507	0,891	0,997

Tabelle 2.2: Geburtstagsproblem

$$= \left(1 - \frac{1}{M}\right)\left(1 - \frac{2}{M}\right) \ldots \left(1 - \frac{n-1}{M}\right)$$

und

$$P(A_n) = 1 - \left(1 - \frac{1}{M}\right) \ldots \left(1 - \frac{n-1}{M}\right)$$

Für manche n gibt Tabelle 2.2 die numerischen Wahrscheinlichkeiten von $P(A_n)$ an.

Es gilt offensichtlich $P(A_n) \approx 1$ für $n \to M$. Interessanterweise ist $P(A_n) \approx 0,5$ für $n = 23$. Dieses Beispiel ist ein Spezialfall eines so genannten *Urnenmodells*: In einer Urne liegen M durchnummerierte Bälle. Aus dieser Urne werden n Stück auf gut Glück mit Zurücklegen entnommen. Wie groß ist die Wahrscheinlichkeit, dass in der Stichprobe mindestens 2 Gleiche vorkommen?

3. *Urnenmodelle*: In einer Urne gibt es M durchnummerierte Bälle. Es werden n Stück "zufällig" entnommen. Das Ergebnis dieses Experimentes ist eine Stichprobe $(j_1, \ldots, j_n)$, wobei j_m die Nummer des Balls in der m–ten Ziehung ist. Es werden folgende Arten der Ziehung betrachtet:

   - *mit Zurücklegen*
   - *ohne Zurücklegen*
   - *mit Reihenfolge*
   - *ohne Reihenfolge*

   Das Geburtstagsproblem ist somit ein Urnenproblem mit Zurücklegen und mit Reihenfolge.

   Demnach werden auch folgende Grundräume betrachtet:

   (a) *Ziehen mit Reihenfolge und mit Zurücklegen:*

$$\Omega = \{\omega = (\omega_1, \ldots, \omega_n) : \omega_i \in \underbrace{\{1, \ldots, M\}}_{=K}\} = K^n, \quad |\Omega| = M^n$$

(b) *Ziehen mit Reihenfolge und ohne Zurücklegen:*

$$\Omega = \{\omega = (\omega_1, \ldots, \omega_n) : \omega_i \in K, \omega_i \neq \omega_j, i \neq j\},$$

$$|\Omega| = M(M-1)\ldots(M-n+1) = \frac{M!}{(M-n)!}$$

Spezialfall: M=n (Permutationen): $\Longrightarrow |\Omega| = M!$

(c) *Ziehen ohne Reihenfolge und mit Zurücklegen:*

$$\Omega = \{\omega = (\omega_1, \ldots, \omega_n) : \omega_i \in K, \omega_1 \leq \omega_2 \leq \cdots \leq \omega_n\}$$

Dies ist äquivalent zu der Verteilung von n Teilchen auf M Zellen ohne Reihenfolge $\Longleftrightarrow$ das Verteilen von $M-1$ Trennwänden der Zellen unter n Teilchen (vgl. Abb. 2.3). Daher ist

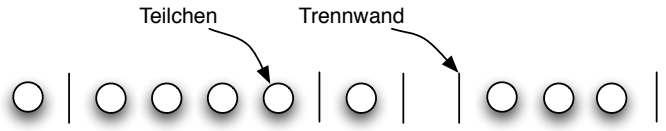


Abbildung 2.3: Ziehen ohne Reihenfolge und mit Zurücklegen

$$|\Omega| = \frac{(M+n-1)!}{n!(M-1)!} = \binom{M+n-1}{n}$$

(d) *Ziehen ohne Reihenfolge und ohne Zurücklegen:*

$$\Omega = \{\omega = (\omega_1, \ldots, \omega_n); \omega_i \in K, \omega_1 < \omega_2 < \cdots < \omega_n\}$$

$$|\Omega| = \frac{M!}{(M-n)!n!} = \binom{M}{n}$$

Ein Experiment der Mehrfachziehung aus einer Urne entspricht der Verteilung von n (unterschiedlichen oder nicht unterscheidbaren) Teilchen (wie z.B. Elektronen, Protonen, usw.) auf M Energieebenen oder Zellen (mit oder ohne Mehrfachbelegung dieser Ebenen) in der statistischen Physik. Die entsprechenden Namen der Modelle sind in Tabelle 2.3 zusammengeführt. So folgen z.B. Elektronen, Protonen und Neutronen der so genannten *Fermi–Dirac–Statistik* (nicht unterscheidbare Teilchen ohne Mehrfachbelegung). Photonen und Prionen folgen der *Bose–Einstein–Statistik* (nicht unterscheidbare Teilchen mit Mehrfachbelegung). Unterscheidbare Teilchen, die dem *Exklusionsprinzip von Pauli* folgen (d.h. ohne Mehrfachbelegung), kommen in der Physik nicht vor.

Auswahl von n aus M Kugeln in einer Urne	**mit Zurücklegen**	**ohne Zurücklegen**	
mit Reihenfolge	M^n (*Maxwell–Boltzmann–Statistik*)	$\frac{M!}{(M-n)!}$	**unterscheidbare Teilchen**
ohne Reihenfolge	$\binom{M+n-1}{n}$ (*Bose–Einstein–Statistik*)	$\binom{M}{n}$ (*Fermi–Dirac–Statistik*)	**nicht unterscheidbare Teilchen**
	mit Mehrfachbelegung	**ohne Mehrfachbelegung**	**Verteilung von n Teilchen auf M Zellen.**

Tabelle 2.3: Die Potenz $|\Omega|$ der Grundgesamtheit Ω in Urnenmodellen.

4. *Lotterie–Beispiele:* ein Urnenmodell ohne Reihenfolge und ohne Zurücklegen; In einer Lotterie gibt es M Lose (durchnummeriert von 1 bis M), davon n Gewinne ($M \geq 2n$). Man kauft n Lose. Mit welcher Wahrscheinlichkeit gewinnt man mindestens einen Preis?
Laut Tabelle 2.3 ist $|\Omega| = \binom{M}{n}$,

$$\Omega = \{\omega = (\omega_1, \ldots, \omega_n) : \omega_i \neq \omega_j, i \neq j, \omega_i \in \{1 \ldots M\}\}\,.$$

Sei $A = \{\text{es gibt mind. 1 Preis}\}$.

$$\begin{aligned} P(A) &= 1 - P(\bar{A}) = 1 - P(\text{es werden keine Preise gewonnen}) \\ n &= 1 - \frac{\binom{M-n}{n}}{|\Omega|} = 1 - \frac{\frac{(M-n)!}{n!(M-2n)!}}{\frac{M!}{n!(M-n)!}} \\ &= 1 - \frac{(M-n)(M-n-1)\ldots(M-2n+1)}{M(M-1)\ldots(M-n+1)} \\ &= 1 - \left(1 - \frac{n}{M}\right)\left(1 - \frac{n}{M-1}\right)\ldots\left(1 - \frac{n}{M-n+1}\right) \end{aligned}$$

Um ein Beispiel zu geben, sei $M = n^2$. Dann gilt:
$P(A) \underset{n\to\infty}{\longrightarrow} 1 - e^{-1} \approx 0,632$, denn $e^x = \lim_{n\to\infty}(1 + \frac{x}{n})^n$. Die Konvergenz ist schnell, $P(A) = 0,670$ schon für $n = 10$.

5. *Hypergeometrische Verteilung:* Nehmen wir jetzt an, dass M Kugeln in der Urne zwei Farben tragen können: schwarz und weiß. Seien S schwarze und W weiße Kugeln gegeben ($M = S + W$). Wie groß ist die Wahrscheinlichkeit, dass aus n zufällig entnommenen Kugeln (ohne Reihenfolge und ohne Zurücklegen) s schwarz sind?
Sei $A = \{$unter n entnommenen Kugeln s schwarze$\}$. Dann ist

$$P(A) = \frac{\binom{S}{s}\binom{W}{n-s}}{\binom{M}{n}}.$$

Diese Wahrscheinlichkeiten bilden die so genannte *hypergeometrische Verteilung*.

Um ein numerisches Beispiel zu geben, seien 36 Spielkarten gegeben. Sie werden zufällig in zwei gleiche Teile aufgeteilt. Wie groß ist die Wahrscheinlichkeit, dass die Anzahl von roten und schwarzen Karten in diesen beiden Teilen gleich ist?
Lösung: hypergeometrische Wahrscheinlichkeiten mit $M = 36$,
$S = W = n = 18, \quad s = \frac{18}{2} = 9, \quad w = s = 9$. Dann ist

$$P(A) = \frac{\binom{18}{9}\binom{18}{9}}{\binom{36}{18}} = \frac{(18!)^4}{36!(9!)^4}.$$

Wenn man die Formel von Stirling

$$n! \approx \sqrt{2\pi n}\, n^n e^{-n}$$

benutzt, so kommt man auf

$$P(A) \approx \frac{(\sqrt{2\pi 18}\cdot 18^{18}e^{-18})^4}{\sqrt{2\pi 36}\cdot 36^{36}e^{-36}(\sqrt{2\pi 9}\cdot 9^9 e^{-9})^4} \approx \frac{2}{\sqrt{18\pi}} \approx \frac{4}{15} \approx 0.26$$

2.3.2 Geometrische Wahrscheinlichkeiten

Hier sei ein Punkt π zufällig auf eine beschränkte Teilmenge Ω von $\mathbb{R}^d$ geworfen. Wie groß ist die Wahrscheinlichkeit, dass π die Teilmenge $A \subset \Omega$ trifft? Um dieses Experiment formalisieren zu können, dürfen wir nur solche Ω und A zulassen, für die der Begriff des d–dimensionalen Volumens (Lebesgue–Maß) wohl definiert ist. Daher werden wir nur Borelsche Teilmengen von $\mathbb{R}^d$ betrachten. Also sei $\Omega \in \mathcal{B}_{\mathbb{R}^d}$ und $|\cdot|$ das Lebesgue–Maß auf $\mathbb{R}^d$, $|\Omega| < \infty$. Sei $\mathcal{F} = \mathcal{B}_{\mathbb{R}^d} \cap \Omega$ (vgl. Abb. 2.4).

Definition 2.3.2

1. Das Wahrscheinlichkeitsmaß auf $(\Omega, \mathcal{F})$ gegeben durch

$$P(A) = \frac{|A|}{|\Omega|}, \quad A \in \mathcal{F}$$

heißt *geometrische Wahrscheinlichkeit* auf Ω.

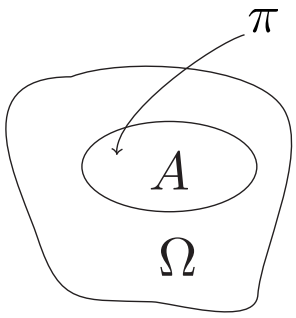


Abbildung 2.4: Zufälliger Punkt π auf Ω.

2. Das Tripel $(\Omega, \mathcal{F}, P)$ heißt *geometrischer Wahrscheinlichkeitsraum.*

Beispiel 2.3.2

Die Koeffizienten p und q einer quadratischen Gleichung $x^2 + px + q = 0$ werden zufällig im Intervall $(0, 1)$ gewählt. Wie groß ist die Wahrscheinlichkeit, dass die Lösungen x_1, x_2 dieser Gleichung reelle Zahlen sind?
Hier ist $\Omega = \{(p, q) : p, q \in (0, 1)\} = (0, 1)^2, \quad \mathcal{F} = \mathcal{B}_{\mathbb{R}^2} \cap \Omega.$

$$A = \{x_1, x_2 \in \Omega\} = \{(p, q) \in \Omega : p^2 \geq 4q\},$$

denn $x_1, x_2 \in \mathbb{R}$ genau dann, wenn die Diskriminante $D = p^2 - 4q \geq 0$. Also gilt $A = \{(p, q) \in [0, 1]^2 : q \leq \frac{1}{4}p^2\}$ und

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\int_0^1 \frac{1}{4}p^2 dp}{1} = \frac{1}{12},$$

vgl. Abb. 2.5.

2.3.3 Bedingte Wahrscheinlichkeiten

Um den Begriff der bedingten Wahrscheinlichkeit intuitiv einführen zu können, betrachten wir zunächst das Beispiel der klassischen Wahrscheinlichkeiten: Sei $(\Omega, \mathcal{F}, P)$ ein Laplacescher Wahrscheinlichkeitsraum mit $|\Omega| = N$. Seien A und B Ereignisse aus $\mathcal{F}$. Dann gilt

$$P(A) = \frac{|A|}{N}, \quad P(A \cap B) = \frac{|A \cap B|}{N}.$$

Wie groß ist die Wahrscheinlichkeit $P(A|B)$ von A unter der Bedingung, dass B eintritt?
Da B eingetreten ist, ist die Gesamtanzahl aller Elementarereignisse hier gleich $|B|$. Die Elementarereignisse, die zu A beim Eintreten von B führen,

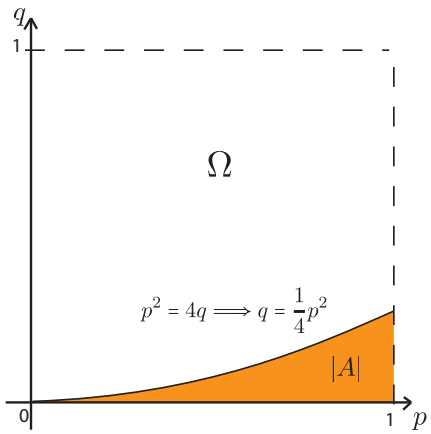


Abbildung 2.5: Wahrscheinlichkeit für reelle Lösungen einer quadratischen Gleichung

liegen alle in $A \cap B$. Somit ist die Anzahl der "günstigen" Fälle hier $|A \cap B|$ und wir bekommen

$$P(A|B) = \frac{|A \cap B|}{|B|} = \frac{|A \cap B|/N}{|B|/N} = \frac{P(A \cap B)}{P(B)}.$$

Dies führt zu folgender Definition:

Definition 2.3.3
Sei $(\Omega, \mathcal{F}, P)$ ein beliebiger Wahrscheinlichkeitsraum, $A, B \in \mathcal{F}$, $P(B) > 0$. Dann ist die *bedingte Wahrscheinlichkeit* von A unter der Bedingung B gegeben durch

$$P(A|B) = \frac{P(A \cap B)}{P(B)}\,.$$

Diese Definition kann in Form des sogenannten Multiplikationssatzes gegeben werden:

$$P(A \cap B) = P(A|B) \cdot P(B)\,.$$

Übungsaufgabe 2.3.1 Zeigen Sie, dass $P(\cdot|B)$ für $B \in \mathcal{F}$, $P(B) > 0$ ein Wahrscheinlichkeitsmaß auf $(\Omega, \mathcal{F})$ ist.

Satz 2.3.1 Seien $A_1, \ldots, A_n \in \mathcal{F}$ Ereignisse mit $P(A_1 \cap \cdots \cap A_{n-1}) > 0$, dann gilt $P(A_1 \cap \cdots \cap A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2) \cdot \cdots \cdot P(A_n|A_1 \cap \cdots \cap A_{n-1})$.

Übungsaufgabe 2.3.2 Beweisen Sie den Satz 2.3.1.
Beweisidee: Induktion bezüglich n.

An dieser Stelle sollte man zu den stochastisch unabhängigen Ereignissen zurückkehren. A und B sind nach Definition 2.2.4 unabhängig, falls $P(A \cap B) = P(A) \cdot P(B)$. Dies ist äquivalent zu $P(A|B) = P(A)$, falls $P(B) > 0$. Es sei allerdings an dieser Stelle angemerkt, dass die Definition 2.2.4 allgemeiner ist, weil sie auch den Fall $P(B) = 0$ zulässt.

Übungsaufgabe 2.3.3 Zeigen Sie folgendes:

1. Seien $A, B \in \mathcal{F}$. A und B sind (stochastisch) unabhängig genau dann, wenn A und $\bar{B}$ oder ($\bar{A}$ und $\bar{B}$) unabhängig sind.

2. Seien $A_1, \ldots, A_n \in \mathcal{F}$. Ereignisse $A_1, \ldots, A_n$ sind stochastisch unabhängig in ihrer Gesamtheit genau dann, wenn $B_1, \ldots, B_n$ unabhängig in ihrer Gesamtheit sind, wobei $B_i = A_i$ oder $B_i = \bar{A}_i$ für $i = 1, \ldots, n$.

3. Seien $A, B_1, B_2 \in \mathcal{F}$ mit $B_1 \cap B_2 = \emptyset$. Sei A und B_1, A und B_2 unabhängig. Zeigen Sie, dass A und $B_1 \cup B_2$ ebenfalls unabhängig sind.

Bemerkung 2.3.2 Der in Definition 2.2.4 gegebene Begriff der stochastischen Unabhängigkeit ist viel allgemeiner als die sogenannte Unabhängigkeit im Sinne des Gesetzes von Ursache und Wirkung. In den folgenden Beispielen wird man sehen, dass zwei Ereignisse stochastisch unabhängig sein können, obwohl ein kausaler Zusammenhang zwischen ihnen besteht. Somit ist die stochastische Unabhängigkeit allgemeiner, und nicht an das Gesetz von Ursache und Wirkung gebunden. In der Praxis allerdings ist man gut beraten, Ereignisse, die keinen kausalen Zusammenhang haben als stochastisch unabhängig zu deklarieren.

Beispiel 2.3.3

1. *Abhängige und unabhängige Ereignisse:*
   Es werde ein Punkt $\pi = (X, Y)$ zufällig auf $[0,1]^2$ geworfen. $\Omega = [0,1]^2, \mathcal{F} = \mathcal{B}_{\mathbb{R}^2} \cap [0,1]$. Betrachten wir $A = \{X \geq a\}$ und $B = \{Y \geq b\}$. Dann gilt

$$\begin{aligned} P(A \cap B) &= P(X \geq a, Y \geq b) \\ &= \frac{(1-a)(1-b)}{1} \\ &= P(A) \cdot P(B)\,, \end{aligned}$$

   insofern sind A und B stochastisch unabhängig. Allerdings kann für $B' = \{\pi \in \Delta CDE\}$ leicht gezeigt werden, dass A und B', B und B' stochastisch abhängig sind, siehe Abbildung 2.6.

2. Es können $n + 1$ Ereignisse konstruiert werden, die abhängig sind, wobei beliebige n von ihnen unabhängig sind, $\forall n \in \mathbb{N}$.

3. *Kausale und stochastische Unabhängigkeit:*
   Auf das Intervall $[0,1]$ wird auf gut Glück ein Punkt π geworfen. Sei

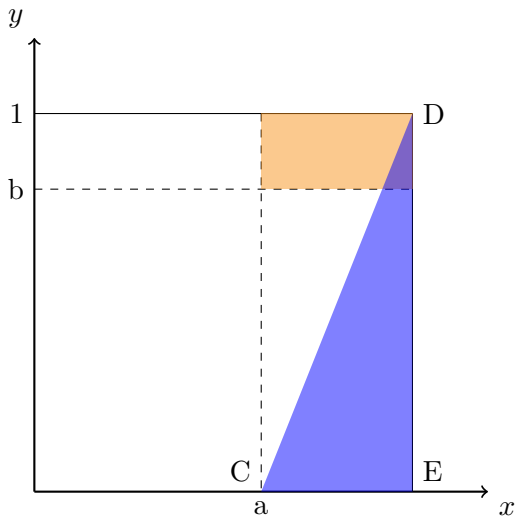


Abbildung 2.6: Beispiel 2.3.3 1.

x die Koordinate von π in $[0,1]$. Betrachten wir die binäre Zerlegung der Zahl x:

$$x = \sum_{k=1}^{\infty} \frac{a_k}{2^k}, \ a_n \in \{0,1\}.$$

Dann ist klar, dass es einen starken kausalen Zusammenhang zwischen $\{a_n\}_{n=1}^{\infty}$ gibt, weil sie alle durch x verbunden sind. Man kann jedoch zeigen, dass die Ereignisse $B_k = \{a_k = j\}, k \in \mathbb{N}$ für alle $j = 0,1$ unabhängig in ihrer Gesamtheit sind, und dass $P(a_k = j) = 1/2 \ \forall k \in \mathbb{N}$, $j = 0,1$.

Definition 2.3.4 Sei $\{B_n\}$ eine endliche oder abzählbare Folge von Ereignissen aus $\mathcal{F}$. Sie heißt eine *messbare Zerlegung* von Ω, falls

1. B_n paarweise disjunkt sind: $B_i \cap B_j = \emptyset, \quad i \neq j$
2. $\bigcup_n B_n = \Omega$
3. $P(B_n) > 0 \quad \forall n.$

Satz 2.3.2 *(Formel der totalen Wahrscheinlichkeit, Bayes'sche Formel):* Sei $\{B_n\} \subset \mathcal{F}$ eine messbare Zerlegung von Ω und $A \in \mathcal{F}$ ein beliebiges Ereignis, dann gilt

1. *Die Formel der totalen Wahrscheinlichkeit:*

$$P(A) = \sum_n P(A|B_n) \cdot P(B_n)$$

2. *Bayes'sche Formel:*

$$P(B_i|A) = \frac{P(B_i) \cdot P(A|B_i)}{\sum_n P(A|B_n) \cdot P(B_n)} \quad \forall i$$

falls $P(A) > 0$. Die Summen in 1) und 2) können endlich oder unendlich sein, je nach Anzahl der B_n.

Beweis 1. Da $\Omega = \bigcup_n B_n$, ist $A = A \cap \Omega = A \cap (\bigcup_n B_n) = \bigcup_n (A \cap B_n)$ eine disjunkte Vereinigung von Ereignissen $A \cap B_n$, und es gilt

$$P(A) = P(\bigcup_n (A \cap B_n)) \underbrace{=}_{\sigma\text{–Add. v. } P} \sum_n P(A \cap B_n) \underbrace{=}_{\text{S. 2.3.1}} \sum_n P(A|B_n)P(B_n)$$

2.

$$P(B_i|A) \underbrace{=}_{\text{Def. 2.3.3}} \frac{P(B_i \cap A)}{P(A)} \underbrace{=}_{\text{S. 2.3.1 u. 2.3.2 1)}} \frac{P(B_i)P(A|B_i)}{\sum_n P(A|B_n)P(B_n)}$$

□

Bemerkung 2.3.3 Die Ereignisse B_n heißen oft "Hypothesen". Dann ist $P(B_n)$ die so genannte *a–priori–Wahrscheinlichkeit von* B_n, also vor dem "Experiment" A. Die Wahrscheinlichkeiten $P(B_n|A)$ werden als Wahrscheinlichkeiten des Auftretens von B_n "nach dem Experiment A" interpretiert. Daher heißen sie auch oft '*a–posteriori–Wahrscheinlichkeiten von* B_n'. Die Formel von Bayes verbindet also die a–posteriori–Wahrscheinlichkeiten mit den a–priori–Wahrscheinlichkeiten.

Beispiel 2.3.4

1. *Routing–Problem:*
   Im Internet muss ein Paket von Rechner S (Sender) auf den Rechner E (Empfänger) übertragen werden. In Abb. 2.7 ist die Geometrie des Computernetzes zwischen S und E schematisch dargestellt, wobei R_1, R_2, R_3 und R_4 (und andere Knoten des Graphen) jeweils andere Rechner sind, die sich an der Übertragung beteiligen können. Wir gehen davon aus, dass die Richtung der weiteren Übertragung des Paketes in den Knoten zufällig aus allen möglichen Knoten gewählt wird (mit gleicher Wahrscheinlichkeit). So ist z.B.

$$P(\underbrace{\text{von S wird Router } R_i \text{ gewählt}}_{=A_i}) = \frac{1}{4} \quad i = 1, \ldots, n\,.$$

   Offensichtlich stellen die Ereignisse A_1, A_2, A_3, A_4 eine messbare Zerlegung von Ω dar. Nach Satz 2.3.2, 1) gilt also für $A = \{$das Paket

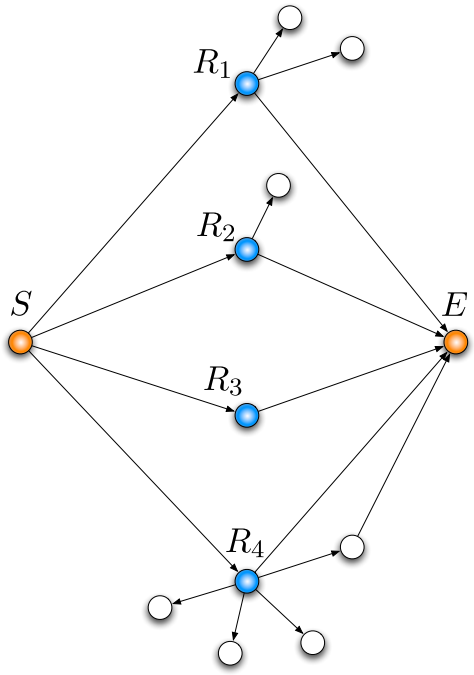


Abbildung 2.7: Routing–Problem: Computernetzwerk

erreicht E aus $S\}$

$$P(A) = \sum_{i=1}^{4} P(A|A_i) \cdot P(A_i) = \frac{1}{4} \sum_{i=1}^{4} P(A|A_i)\,.$$

Dabei können $P(A|A_i)$ aus dem Graphen eindeutig bestimmt werden:

$$P(A|A_1) = \frac{1}{3}\,, \quad P(A|A_2) = \frac{1}{2}\,,$$

$$P(A|A_3) = 1\,, \quad P(A|A_4) = \frac{2}{5}\,.$$

Es gilt also

$$P(A) = \frac{1}{4}\left(\frac{1}{3} + \frac{1}{2} + 1 + \frac{2}{5}\right) = \frac{67}{120} \approx 0,5\,.$$

2. In einer Urne gibt es zwei Münzen. Die erste ist fair (Wahrscheinlichkeit des Kopfes und der Zahl $= \frac{1}{2}$), die zweite ist allerdings nicht fair mit $P(\text{Kopf}) = \frac{1}{3}$. Aus der Urne wird eine Münze zufällig genommen und geworfen. In diesem Wurf ist das Ereignis Kopf. Wie groß ist die

Wahrscheinlichkeit, dass die Münze fair war?
Sei

$$A_1 = \{\text{Faire Münze ausgewählt}\}$$

$$A_2 = \{\text{Nicht faire Münze ausgewählt}\}$$

$$A = \{\text{Es kommt Kopf im Münzwurf}\}$$

$$P(A_1|A) = ?$$

Dann gilt $P(A_1) = P(A_2) = \frac{1}{2}$, $P(A|A_1) = \frac{1}{2}$, $\quad P(A|A_2) = \frac{1}{3}$, daher gilt nach der Bayesschen Formel

$$P(A_1|A) = \frac{P(A_1) \cdot P(A|A_1)}{P(A_1) \cdot P(A|A_1) + P(A_2) \cdot P(A|A_2)} = \frac{\frac{1}{2} \cdot \frac{1}{2}}{\frac{1}{2} \cdot (\frac{1}{2} + \frac{1}{3})} = \frac{3}{5}.$$

Kapitel 3

Zufallsvariablen

3.1 Definition und Beispiele

Definition 3.1.1 1. Eine Abbildung $X : \Omega \to \mathbb{R}$ heißt *Zufallsvariable*, falls sie $\mathcal{B}_{\mathbb{R}}$–messbar ist, mit anderen Worten,

$$\forall B \in \mathcal{B}_{\mathbb{R}} \quad X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}\,.$$

2. Eine Abbildung $X : \Omega \to \mathbb{R}^n$, $n \geq 1$ heißt *Zufallsvektor*, falls sie $\mathcal{B}_{\mathbb{R}^n}$–messbar ist, mit anderen Worten,

$$\forall B \in \mathcal{B}_{\mathbb{R}^n} \quad X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}\,.$$

Offensichtlich bekommt man aus Definition 3.1.1, 2) auch 3.1.1, 1) für $n = 1$.

Beispiel 3.1.1 1. *Indikator–Funktion eines Ereignisses:*
Sei $(\Omega, \mathcal{F}, P)$ ein Wahrscheinlichkeitsraum und A ein Ereignis aus $\mathcal{F}$. Betrachten wir

$$X(\omega) = I_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \omega \notin A \end{cases}.$$

Diese Funktion von ω nennt man *Indikator–Funktion des Ereignisses* A. Sie ist offensichtlich messbar und somit eine Zufallsvariable:

$$X^{-1}(B) = \left.\begin{cases} A & \text{falls} \quad 1 \in B, 0 \notin B \\ \bar{A} & \text{falls} \quad 1 \notin B, 0 \in B \\ \Omega & \text{falls} \quad 0, 1 \in B \\ \emptyset & \text{falls} \quad 0, 1 \notin B \end{cases}\right| \in \mathcal{F} \quad \forall B \in \mathcal{B}_{\mathbb{R}}$$

2. *n–maliger Münzwurf:*
Sei $\Omega = \{\omega = (\omega_1, \ldots, \omega_n) : \omega_i \in \{0, 1\}\} = \{0, 1\}^n$ mit

$$\omega_i = \begin{cases} 1, & \text{falls Kopf im i–ten Münzwurf} \\ 0, & \text{sonst} \end{cases}$$

für $i = 1, \ldots, n$. Sei $\mathcal{F} = \mathcal{P}(\Omega)$. Definieren wir

$$X(\omega) = X((\omega_1, \ldots, \omega_n)) = \sum_{i=1}^{n} \omega_i$$

als die Anzahl der Köpfe im n–maligen Münzwurf, so kann man zeigen, dass X $\mathcal{F}$-messbar ist und somit eine Zufallsvariable ist.

Satz 3.1.1 Eine Abbildung $X : \Omega \to \mathbb{R}$ ist genau dann eine Zufallsvariable, wenn $\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F} \quad \forall x \in \mathbb{R}$.

3.2 Verteilungsfunktion

Definition 3.2.1 Sei $(\Omega, \mathcal{F}, P)$ ein beliebiger Wahrscheinlichkeitsraum und $X : \Omega \to \mathbb{R}$ eine Zufallsgröße.

1. Die Funktion $F_X(x) = P(\{\omega \in \Omega : X(\omega) \leq x\}), \quad x \in \mathbb{R}$ heißt *Verteilungsfunktion* von X. Offensichtlich ist $F_X : \mathbb{R} \to [0, 1]$.

2. Die Mengenfunktion $P_X : \mathcal{B}_{\mathbb{R}} \to [0, 1]$ gegeben durch

$$P_X(B) = P(\{\omega \in \Omega : X(\omega) \in B\}), B \in \mathcal{B}_{\mathbb{R}}$$

heißt *Verteilung* von X.

Bemerkung 3.2.1 Folgende gekürzte Schreibweise wird benutzt:

$$F_X(x) = P(X \leq x), \quad P_X(B) = P(X \in B).$$

Beispiel 3.2.1
Hier geben wir Verteilungsfunktionen für Zufallsvariablen aus dem Beispiel 3.1.1 an.

1. *Indikator–Funktion:*
Sei $X(\omega) = I_A(\omega)$. Dann ist

$$F_X(x) = P(I_A \leq x) = \begin{cases} 1, & x \geq 1, \\ P(\overline{A}), & x \in [0, 1), \\ 0, & x < 0 \end{cases}$$

vgl. Abb. 3.1.

2. *n–maliger Münzwurf:*
Sei X = Anzahl Kopf in n Münzwürfen. $P(\text{Kopf in einem Wurf}) = p$, $p \in (0, 1)$. Dann gilt

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, \ldots, n,$$

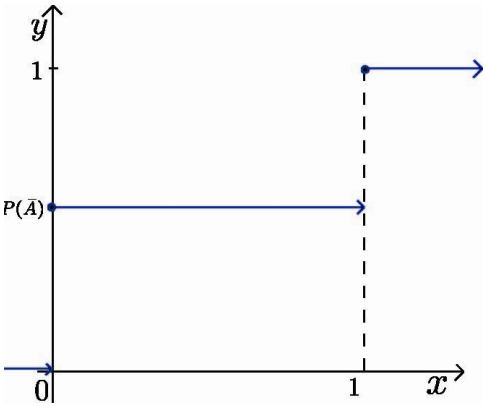


Abbildung 3.1: Verteilungsfunktion von I_A

und somit

$$F_X(x) = P(X \leq x) = \sum_{0 \leq k \leq [x]} P(x = k) = \sum_{k=0}^{[x]} \binom{n}{k} p^k (1-p)^{n-k} ,$$

$\forall\, x \in [0, n]$, vgl. Abb. 3.2. Es gilt

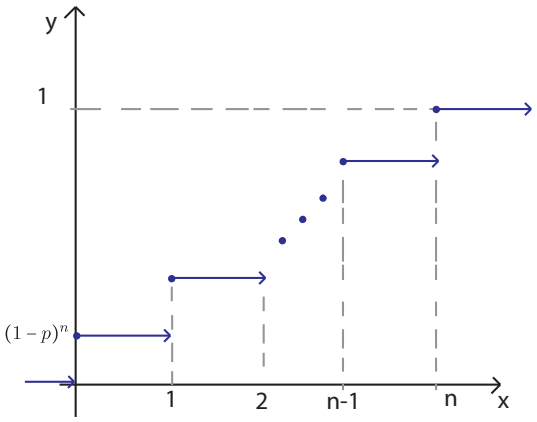


Abbildung 3.2: Verteilungsfunktion einer $Bin(n, p)$–Verteilung

$$\begin{aligned} F_X(0) &= P(X \leq 0) = P(X = 0) = (1-p)^n , \\ F_X(x) &= P(X \leq x) = 0 \quad \text{für } x < 0 , \\ F_X(n) &= P(X \leq n) = 1 . \end{aligned}$$

Diese Verteilung wird später *Binomial–Verteilung* mit Parametern n, p genannt: $Bin(n, p)$

Satz 3.2.1 Sei X eine beliebige Zufallsvariable und $F_X : \mathbb{R} \to [0,1]$ ihre Verteilungsfunktion. F_X besitzt folgende Eigenschaften:

1. *Asymptotik:* $\lim_{x\to-\infty} F_X(x) = 0, \quad \lim_{x\to+\infty} F_X(x) = 1.$
2. *Monotonie:* $F_X(x) \leq F_X(x+h), \quad \forall x \in \mathbb{R},\, h \geq 0\,.$
3. *Rechtsseitige Stetigkeit:* $\lim_{x\to x_0+0} F_X(x) = F_X(x_0) \quad \forall x_0 \in \mathbb{R}\,.$

Bemerkung 3.2.2

1. Im Satz 3.2.1 wurde gezeigt, dass eine Verteilungsfunktion F_X monoton nicht–fallend, rechtsseitig stetig und beschränkt auf $[0,1]$ ist. Diese Eigenschaften garantieren, dass F_X höchstens abzählbar viele Sprungstellen haben kann. In der Tat kann F_X wegen $F_X \uparrow$ und $0 \leq F_X \leq 1$ nur eine endliche Anzahl von Sprungstellen mit Sprunghöhe $> \varepsilon$ besitzen, $\forall \varepsilon > 0$. Falls ε_n die Menge $\mathbb{Q}$ aller rationaler Zahlen durchläuft, wird somit gezeigt, dass die Anzahl aller möglichen Sprungstellen höchstens abzählbar sein kann. Die Grafik einer typischen Verteilungsfunktion ist in Abb. 3.3 dargestellt.

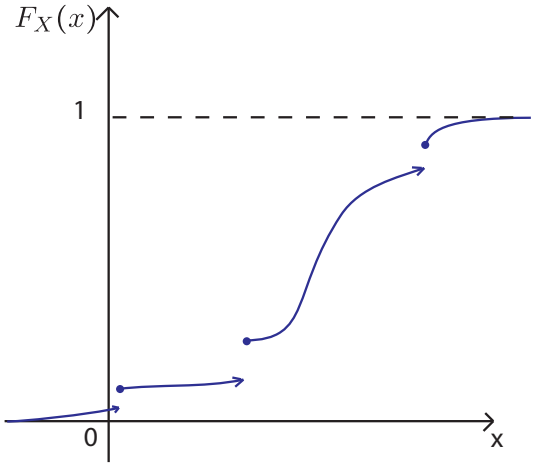


Abbildung 3.3: Typische Verteilungsfunktion

2. Mit Hilfe von F_X können folgende Wahrscheinlichkeiten leicht berechnet werden: $\forall a, b$ gilt: $-\infty \leq a < b \leq +\infty$

$$P(a < X \leq b) = F_X(b) - F_X(a)\,,$$
$$P(a \leq X \leq b) = F_X(b) - \lim_{x\to a-o} F_X(x)\,,$$

denn

$$P(a < X \leq b) = P(\{X \leq b\} \backslash \{X \leq a\}) = P(X \leq b) - P(X \leq a)$$

$$= F_X(b) - F_X(a)\,,$$
$$P(a \leq X \leq b) = P(X \leq b) - P(X < a) = F_X(b) - \lim_{x \to a-0} F_X(x)$$

mit $P(X < a) = P(X \leq a) - P(X = a) = \lim_{x \to a-0} F_X(x)$ nach Stetigkeit von P_X.

Da $P(X < a) = F(X \leq a) - P(X = a)$ gilt, ist somit

$$\lim_{x \to a-0} F_X(x) \neq F_X(a)$$

und F_X im Allgemeinen nicht linksseitig stetig.

Übungsaufgabe 3.2.1 Drücken Sie die Wahrscheinlichkeiten $P(a < X < b)$ und $P(a \leq X < b)$ mit Hilfe von F_X aus.

Satz 3.2.2 Falls eine Funktion $F(x)$ die Eigenschaften 1) bis 3) des Satzes 3.2.1 erfüllt, dann existiert ein Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, P)$ und eine Zufallsvariable X, definiert auf diesem Wahrscheinlichkeitsraum, derart, dass $F_X(x) = F(x)$, $\forall x \in \mathbb{R}$.

Satz 3.2.3 Die Verteilung P_X einer Zufallsvariable X wird eindeutig durch die Verteilungsfunktion F_X von X bestimmt.

3.3 Grundlegende Klassen von Verteilungen

In diesem Abschnitt werden wir Grundtypen von Verteilungen betrachten, die dem Schema aus Abbildung 3.4 zu entnehmen sind.

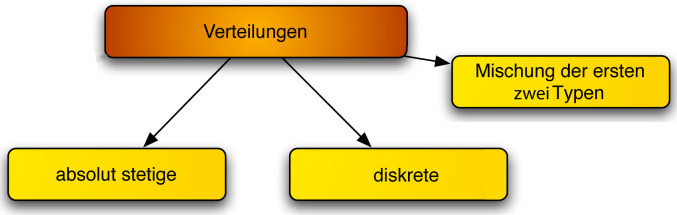


Abbildung 3.4: Verteilungstypen

3.3.1 Diskrete Verteilungen

Definition 3.3.1 1. Die Verteilung einer Zufallsvariablen X heißt *diskret*, falls eine höchstens abzählbare Teilmenge $C \subset \mathbb{R}$ (Wertebereich von X) mit $P(X \in C) = 1$ existiert. Manchmal wird auch die Zufallsvariable X selbst als diskret bezeichnet.

2. Falls X eine diskrete Zufallsvariable mit Wertebereich $C = \{x_1, x_2, x_3, \dots\}$ ist, dann heißt $\{p_k\}$ mit $p_k = P(X = x_k), \quad k = 1, 2, \dots$ *Wahrscheinlichkeitsfunktion* oder *Zähldichte* von X.

Bemerkung 3.3.1

1. Beispiele für diskrete Wertebereiche C sind $\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \{0, 1, \dots, n\}$, $n \in \mathbb{N}$.

2. Für die Zähldichte $\{p_k\}$ einer diskreten Zufallsvariable X gilt offenbar $0 \leq p_k \leq 1 \quad \forall n$ und $\sum_k p_k = 1$. Diese Eigenschaften sind für eine Zähldichte charakteristisch.

3. Die Verteilung P_X einer diskreten Zufallsvariable X wird eindeutig durch ihre Zähldichte $\{p_k\}$ festgelegt:

$$P_X(B) = P_X(B \cap C) = P_X\Big(\bigcup_{x_i \in B} \{x_i\}\Big) = \sum_{x_i \in B} P(X = x_i)$$
$$= \sum_{x_i \in B} p_i, \; B \in \mathcal{B}_{\mathbb{R}}\,.$$

   Insbesondere gilt $F_X(x) = \sum_{x_k \leq x} p_k \implies P_X$ festgelegt nach Satz 3.2.3.

Wichtige diskrete Verteilungen:

Die Beispiele 3.1.1 und 3.2.1 liefern uns zwei wichtige diskrete Verteilungen mit Wertebereichen $\{0, 1\}$ und $\{0, 1, \dots, n\}$. Das sind

1. *Bernoulli–Verteilung:*
   $X \sim Ber(p), \quad p \in [0, 1]$ (abkürzende Schreibweise für "Zufallsvariable X ist Bernoulli–verteilt mit Parameter p"), falls

$$X = \begin{cases} 1, & \text{mit Wahrscheinlichkeit p,} \\ 0, & \text{mit Wahrscheinlichkeit 1-p}\,. \end{cases}$$

   Dann gilt $C = \{0, 1\}$ und $p_0 = 1 - p, \quad p_1 = p$ (vgl. Beispiel 3.1.1, 1) mit $X = I_A$).

2. *Binomialverteilung:*
   $X \sim Bin(n, p)$, $p \in [0, 1], n \in \mathbb{N}$, falls $C = \{0, \dots, n\}$ und

$$P(X = k) = p_k = \binom{n}{k} \cdot p^k (1 - p)^{n-k}, \quad k = 0, \dots, n\,.$$

   Interpretation:
   $X = \#\{$ Erfolge in einem n mal unabhängig wiederholten Versuch$\}$, wobei $p =$ Erfolgswahrscheinlichkeit in einem Versuch (vgl. Beispiel 3.2.1, 2) mit $X = \#\{\text{Kopf}\}$).

3. *Geometrische Verteilung:*
   $X \sim Geo(p)$, $p \in [0,1]$, falls $C = \mathbb{N}$, und

$$P(X = k) = p_k = (1-p)p^{k-1}, \quad k \in \mathbb{N}\,.$$

   Interpretation: $X = \#\{\text{unabhängige Versuche bis zum ersten Erfolg}\}$, wobei $1-p =$ Erfolgswahrscheinlichkeit in einem Versuch.

4. *Hypergeometrische Verteilung:*
   $X \sim HG(M,S,n)$, $M,S,n \in \mathbb{N}, S,n \leq M$, falls

$$X : \Omega \to \{0,1,2,\ldots,\min\{n,S\}\}$$

   und

$$P(X = k) = p_k = \frac{\binom{S}{k}\binom{M-S}{n-k}}{\binom{M}{n}}, \quad k = 0,1,\ldots,\min\{n,S\}\,.$$

   Interpretation: Urnenmodell aus Beispiel 2.3.1, 5) mit

$$X = \#\{\text{schwarze Kugeln bei } n \text{ Entnahmen aus einer Urne}\}$$

   mit insgesamt S schwarzen und $M-S$ weißen Kugeln.

5. *Gleichverteilung:*
   $X \sim U\{x_1,\ldots,x_n\}$, $n \in \mathbb{N}$, falls $X : \Omega \to \{x_1,\ldots,x_n\}$ mit

$$p_k = P(X = x_k) = \frac{1}{n}, \quad k = 1,\ldots,n$$

   (wobei U in der Bezeichnung von Englischen “uniform” kommt).

   Interpretation: $\{p_k\}$ ist eine Laplacesche Verteilung (klassische Definition von Wahrscheinlichkeiten, vgl. Abschnitt 2.3.1).

6. *Poisson–Verteilung:*
   $X \sim Poisson(\lambda)$, $\lambda > 0$, falls $X : \Omega \to \{0,1,2,\ldots\} = \mathbb{N} \cup \{0\}$ mit

$$p_k = P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k \in \mathbb{N} \cup \{0\}\,.$$

   Interpretation: $X = \#\{\text{Ereignisse im Zeitraum } [0,1]\}$, λ ist die Rate (Häufigkeit), mit der Ereignisse passieren können, wobei

$$P(1 \text{ Ereignis tritt während } \Delta t \text{ ein}) = \lambda|\Delta t| + o(|\Delta t|)\,,$$

$$P(>1 \text{ Ereignis tritt während } \Delta t \text{ ein}) = o(|\Delta t|), \quad |\Delta t| \to o$$

   und $\#\{\text{Ereignisse in Zeitintervall } \Delta t_i\}$, $i = 1,\ldots,n$ sind unabhängig, falls $\Delta t_i\,, i = 1,\ldots,n$ disjunkte Intervalle aus $\mathbb{R}$ sind. Hier $|\Delta t|$ ist die

Länge des Intervalls Δt.
z. B.

$$X = \#\{\text{ Schäden eines Versicherers in einem Geschäftsjahr}\}$$

$$X = \#\{\text{ Kundenanrufe eines Festnetzanbieters an einem Tag}\}$$

$$X = \#\{\text{ Elementarteilchen in einem Geiger–Zähler in einer Sekunde}\}\,.$$

Satz 3.3.1 (Approximationssatz)

1. *Binomiale Approximation:*

   Die hypergeometrische Verteilung $HG(M, S, n)$ kann für $M, S \to \infty$, $\frac{S}{M} \to p$ durch eine $Bin(n, p)$–Verteilung approximiert werden: Für $X \sim HG(M, S, n)$ gilt

   $$p_k = P(X = k) = \frac{\binom{S}{k}\binom{M-S}{n-k}}{\binom{M}{n}} \underbrace{\longrightarrow}_{M,S\to\infty,\,\frac{S}{M}\to p} \binom{n}{k} p^k (1-p)^{n-k}$$

   für $k = 1, \ldots, n$.

2. *Poissonsche Approximation oder Gesetz der seltenen Ereignisse:*

   Die Binomialverteilung $Bin(n, p)$ kann für $n \to \infty, p \to 0, np \to \lambda$ durch eine Poisson–Verteilung $Poisson(\lambda)$ approximiert werden: Es sei $X \sim Bin(n, p)$. Dann gilt

   $$p_k = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \underbrace{\longrightarrow}_{n\to\infty,p\to 0,np\to\lambda} e^{-\lambda}\frac{\lambda^k}{k!}$$

   mit $k = 0, 1, 2, \ldots$

Beweis 1. Falls $M, S \to \infty$, $\frac{S}{M} \to p \in (0, 1)$, dann gilt

$$\begin{aligned}
\frac{\binom{S}{k}\binom{M-S}{n-k}}{\binom{M}{n}} &= \frac{\frac{S!}{k!(S-k)!} \cdot \frac{(M-S)!}{(M-S-n+k)!(n-k)!}}{\frac{M!}{n!(M-n)!}} \\
&= \frac{n!}{(n-k)!k!} \underbrace{\frac{S}{M}}_{\to p} \underbrace{\frac{(S-1)}{(M-1)}}_{\to p} \frac{\ldots}{\ldots} \underbrace{\frac{(S-k+1)}{(M-k+1)}}_{\to p} \cdot \\
&\times \underbrace{\frac{(M-S)}{(M-k)}}_{\to 1-p} \underbrace{\frac{(M-S-1)}{\ldots}}_{\to 1-p} \frac{\ldots}{\ldots} \underbrace{\frac{(M-S-n+k+1)}{(M-n+1)}}_{\to 1-p} \\
&\underbrace{\longrightarrow}_{M,S\to\infty\ \ \frac{S}{M}\to p} \binom{n}{k} p^k (1-p)^{n-k}
\end{aligned}$$

2. Falls $n \to \infty$, $p \to 0$, $np \to \lambda > 0$, dann gilt

$$\begin{aligned}\binom{n}{k} p^k (1-p)^{n-k} &= \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\ &= \frac{1}{k!} \underbrace{\frac{n(n-1)\dots(n-k+1)}{n^k}}_{\to 1} \cdot \underbrace{(np)^k}_{\to \lambda^k} \underbrace{\frac{(1-p)^n}{(1-p)^k}}_{\longrightarrow e^{-\lambda}} \\ &\longrightarrow e^{-\lambda} \frac{\lambda^k}{k!} \text{ für } n \to \infty,\ p \to 0,\ np \to \lambda\,,\end{aligned}$$

weil

$$\frac{(1-p)^n}{(1-p)^k} \underset{n\to\infty}{\sim} \frac{(1-\frac{\lambda}{n})^n}{1} \underset{n\to\infty}{\longrightarrow} e^{-\lambda},\ \text{da } p \sim \frac{\lambda}{n}\ (n \to \infty)\,.$$

□

Bemerkung 3.3.2

1. Die Aussage 1) aus Satz 3.3.1 wird dann verwendet, wenn M und S in $HG(M, S, n)$–Verteilung groß werden ($n < 0,1 \cdot M$). Dabei wird die direkte Berechnung von hypergeometrischen Wahrscheinlichkeiten umständlich.

2. Genauso wird die Poisson–Approximation verwendet, falls n groß und p entweder bei 0 oder bei 1 liegt. Dann können binomiale Wahrscheinlichkeiten nur schwer berechnet werden.

3. Bei allen diskreten Verteilungen ist die zugehörige Verteilungsfunktion eine stückweise konstante Treppenfunktion (vgl. Bsp. 1, 2 im Abschnitt 3.2.1).

3.3.2 Absolut stetige Verteilungen

Im Gegensatz zu diskreten Zufallsvariablen ist der Wertebereich einer absolut stetigen Zufallsvariablen überabzählbar.

Definition 3.3.2 Die Verteilung einer Zufallsvariablen X heißt *absolut stetig*, falls die Verteilungsfunktion von F_X folgende Darstellung besitzt:

$$F_X(x) = \int_{-\infty}^{x} f_X(y)\, dy, \quad x \in \mathbb{R}, \tag{3.1}$$

wobei $f_X : \mathbb{R} \to \mathbb{R}_+ = [0, \infty)$ eine Lebesgue–integrierbare Funktion auf $\mathbb{R}$ ist, die *Dichte* der Verteilung von X heißt und das Integral in (3.1) als Lebesgue–Integral zu verstehen ist.
Daher wird oft abkürzend gesagt, dass die Zufallsvariable X absolut stetig (verteilt) mit Dichte f_X ist.

Im folgenden Satz zeigen wir, dass die Verteilung P_X einer absolut stetigen Zufallsvariablen eindeutig durch ihre Dichte f_X bestimmt wird:

Satz 3.3.2 Sei X eine Zufallsvariable mit Verteilung P_X.

1. X ist absolut stetig verteilt genau dann, wenn
$$P_X(B) = \int_B f_X(y)\,dy, \quad B \in \mathcal{B}_{\mathbb{R}}. \tag{3.2}$$

2. Seien X und Y absolut stetige Zufallsvariablen mit Dichten f_X, f_Y und Verteilungen P_X und P_Y. Es gilt $P_X = P_Y$ genau dann, wenn $f_X(x) = f_Y(x)$ für fast alle $x \in \mathbb{R}$, d.h. für alle $x \in \mathbb{R}\backslash A$, wobei $A \in \mathcal{B}_{\mathbb{R}}$ und $\int_A dy = 0$ (das Lebesgue–Maß von A ist Null).

Bemerkung 3.3.3 *(Eigenschaften der absolut stetigen Verteilungen):* Sei X absolut stetig verteilt mit Verteilungsfunktion F_X und Dichte f_X.

1. Für die Dichte f_X gilt: $f_X(x) \geq 0 \quad \forall x$ und $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$ (vgl. Abb. 3.5).
   Diese Eigenschaften sind charakteristisch für eine Dichte, d.h. eine

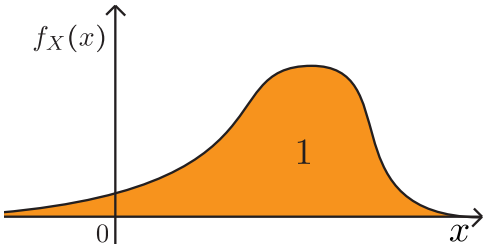


Abbildung 3.5: Die Fläche unter dem Graphen einer Dichtenfunktion ist gleich eins.

   beliebige Funktion f, die diese Eigenschaften erfüllt, ist die Dichte einer absolut stetigen Verteilung.

2. Es folgt aus (3.2), dass
   (a) $P(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(y)\,dy, \quad \forall a < b,\ a, b \in \mathbb{R}$
   (b) $P(X = x) = \int_{\{x\}} f_X(y)\,dy = 0, \quad \forall x \in \mathbb{R},$
   (c) $f_X(x)\Delta x$ als Wahrscheinlichkeit $P(X \in [x, x + \Delta x])$ interpretiert werden kann, falls f_X stetig in der Umgebung von x und Δx klein ist.
   In der Tat, mit Hilfe des Mittelwertsatzes bekommt man
$$P(X \in [x, x + \Delta x]) \quad = \quad \int_x^{x+\Delta x} f_X(y)\,dy$$

$$\begin{aligned} &= && f_X(\xi)\cdot\Delta x, \quad \xi\in(x,x+\Delta x)\\ &\underbrace{=}_{\Delta x\to 0} && (f_X(x)+o(1))\Delta x\\ &= && f_X(x)\cdot\Delta x + o(\Delta x)\,, \end{aligned}$$

weil $\xi \to x$ für $\Delta x \to 0$ und f_X stetig in der Umgebung von x ist.

3. Es folgt aus 2b, dass die Verteilungsfunktion F_X von X eine stetige Funktion ist. F_X kann keine Sprünge haben, weil die Höhe eines Sprunges von F_X in x genau $P(X = x) = 0$ darstellt (vgl. Abb. 3.6).

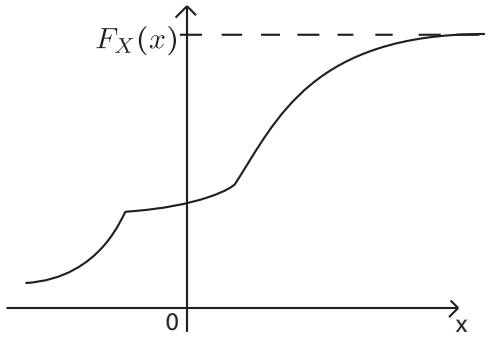


Abbildung 3.6: Eine absolut stetige Verteilungsfunktion

4. Sehr oft wird f_X als (stückweise) stetig angenommen. Dann ist das Integral in Definition 3.3.2 das (uneigentliche) Riemann–Integral. F_X ist im Allgemeinen nur an jeder Stetigkeitsstelle von ihrer Dichte f_X differenzierbar.

5. In den Anwendungen sind Wertebereiche aller Zufallsvariablen endlich. Somit könnte man meinen, dass für Modellierungszwecke nur diskrete Zufallsvariablen genügen. Falls der Wertebereich einer Zufallsvariable X jedoch sehr viele Elemente x enthält, ist die Beschreibung dieser Zufallsvariable mit einer absolut stetigen Verteilung günstiger, denn man braucht nur eine Funktion f_X (Dichte) anzugeben, statt sehr viele Einzelwahrscheinlichkeiten $p_k = P(X = x_k)$ aus den Daten zu schätzen.

Wichtige absolut stetige Verteilungen

1. *Normalverteilung (Gauß–Verteilung):*
   $X \sim N(\mu, \sigma^2)$ für $\mu \in \mathbb{R}$ und $\sigma^2 > 0$, falls

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}$$

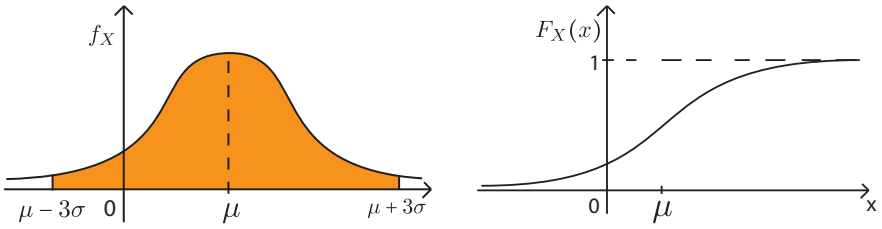


Abbildung 3.7: Dichte und Verteilungsfunktion der $N(\mu, \sigma^2)$–Verteilung

(vgl. Abb. 3.7).
μ heißt der *Mittelwert* von X und σ die *Standardabweichung* bzw. *Streuung*, denn es gilt die sogenannte "3σ–Regel" (Gauß, 1821):

$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \geq 0,9973$$

.

Spezialfall $N(0,1)$: In diesem Fall sieht die Dichte f_X folgendermaßen aus:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in \mathbb{R}\,.$$

Interpretation:
X = Messfehler einer physikalischen Größe μ, σ =Streuung des Messfehlers. Die Verteilungsfunktion $F_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{(y-\mu)^2}{2\sigma^2}}\, dy$ kann nicht analytisch berechnet werden (vgl. Abb. 3.7).

2. *Gleichverteilung auf* $[a,b]$*:*
$X \sim U[a,b]$, $\quad a < b$, $a, b \in \mathbb{R}$, falls

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & x \in [a,b]\,, \\ 0, & \text{sonst} \end{cases} \quad \text{(vgl. Abb. 3.8).}$$

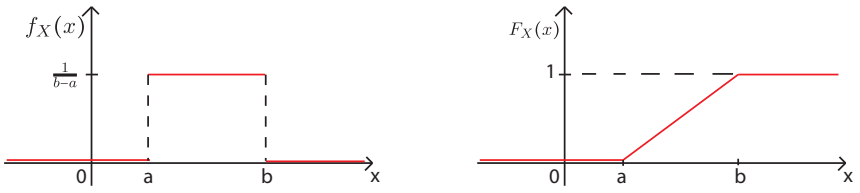


Abbildung 3.8: Dichte und Verteilungsfunktion der Gleichverteilung $U[a,b]$.

Interpretation:
X = Koordinate eines zufällig auf $[a, b]$ geworfenen Punktes (geometrische Wahrscheinlichkeit). Für $F_X(x)$ gilt:

$$F_X(x) = \begin{cases} 1, & x \geq b\,, \\ \frac{x-a}{b-a}, & x \in [a, b)\,, \\ 0, & x < a \ \text{(vgl. Abb. 3.8)}. \end{cases}$$

3. *Exponentialverteilung:*
$X \sim Exp(\lambda)$ für $\lambda > 0$, falls

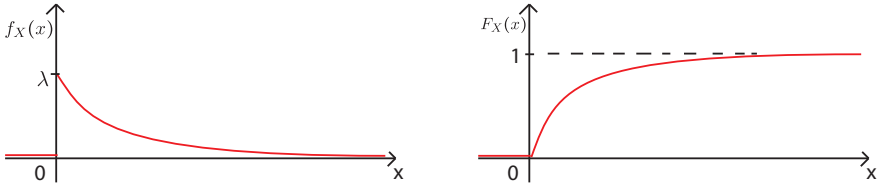


Abbildung 3.9: Dichte und Verteilungsfunktion der Exponentialverteilung $Exp(\lambda)$.

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & \text{sonst (vgl. Abb. 3.9)}. \end{cases}$$

Interpretation:
X = Zeitspanne der fehlerfreien Arbeit eines Geräts, z.B. eines Netzservers oder einer Glühbirne, λ= Alterungsrate des Geräts. $F_X(x)$ hat folgende Gestalt:

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0\,, \\ 0, & x < 0 \ \text{(vgl. Abb. 3.9)}. \end{cases}$$

4. *Cauchy–Verteilung:*
$X \sim Cauchy(\alpha, \lambda)$, falls für $\lambda > 0$, $\alpha \in \mathbb{R}$

$$f_X(x) = \frac{\lambda}{\pi(\lambda^2 + (x-\alpha)^2)}, \quad x \in \mathbb{R}\,, \ \text{vgl. Abb. 3.10}$$

Die Verteilungsfunktion der Cauchy-Verteilung ist dabei

$$F_X(x) = \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{x-\alpha}{\lambda}\right), \quad x \in \mathbb{R}.$$

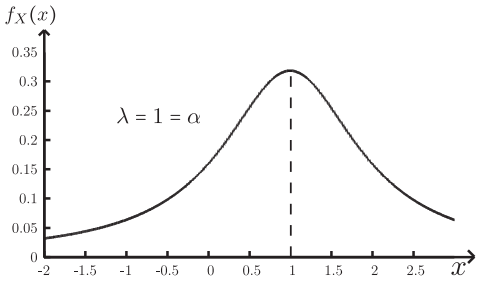


Abbildung 3.10: Dichte der $Cauchy(\alpha, \lambda)$–Verteilung

Interpretation:
Diese Verteilung beschreibt z.B. die Positionen der radioaktiven Teilchen in einem Detektor, sowie die Energie der instabilen Zustände in Kernspaltungsreaktionen (Gesetz von Lorenz).

5. *Pareto–Verteilung:*
   $X \sim \text{Pareto}(\alpha, \mu)$, $\alpha, \mu > 0$, falls

$$f_X(x) = \frac{\alpha \mu^\alpha}{x^{\alpha+1}} I_{[\mu,\infty)}(x), \quad F_X(x) = \left(1 - \frac{\mu^\alpha}{x^\alpha}\right) I_{[\mu,\infty)}(x).$$

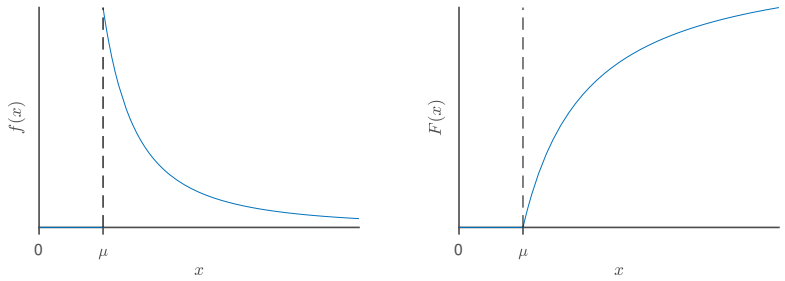


Abbildung 3.11: Dichte und Verteilungsfunktion der $\text{Pareto}(\alpha, \mu)$–Verteilung.

Interpretation:
X = Schadenshöhe einer Police eines Feuerversicherers. Da $P(X > x) = (\mu/x)^\alpha$, $x \to \infty$, nur langsam gegen Null geht (verglichen mit der $N(0,1)$-Verteilung), spricht man hier von einer Verteilung mit *schwerem Tailverhalten* oder von einem *gefährlichen Risiko* X.

6. *Fréchet–Verteilung:*
   $X \sim \text{Fréchet}(\mu, \sigma, \alpha)$, $\alpha, \sigma > 0$, $\mu \in \mathbb{R}$, falls

$$f_X(x) = \alpha \sigma^\alpha (x - \mu)^{-\alpha} e^{-\left(\frac{x-\mu}{\sigma}\right)^{-\alpha}} I_{[\mu,\infty)}(x).$$

Damit ist die Verteilungsfunktion

$$F_X(x) = e^{-\left(\frac{x-\mu}{\sigma}\right)^{-\alpha}} I_{[\mu,\infty)}(x).$$

Die Standard-Fréchet-verteilung hat Parameterwerte $\sigma = 1$, $\mu = 0$:

$$F_X(x) = e^{-x^{-\alpha}} I_{[\mu,\infty)}(x).$$

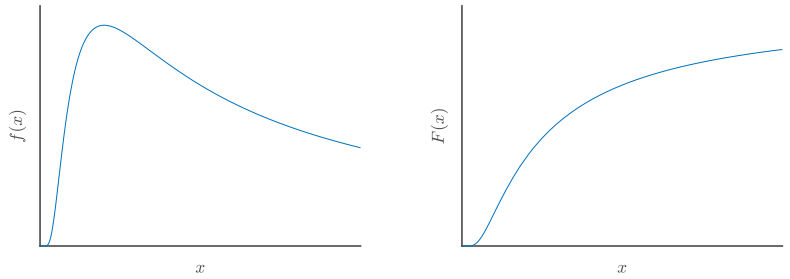


Abbildung 3.12: Dichte- und Verteilungsfunktion der Standard-Fréchet-Verteilung.

Interpretation:
X approximiert normiertes Maximum von n unabhängigen Beobachtungen, die Cauchy- oder Pareto-verteilt sind. Es ist eine der drei möglichen Extremwertverteilungen, zusammen mit der Gumbel- und Weibull-Verteilung.

3.3.3 Mischungen von Verteilungen

Definition 3.3.3 Sei $\{F_n\}_{n=1}^{\infty}$ eine Folge von Verteilungsfunktionen und sei $\{p_n\}_{n=1}^{\infty}$ eine Zähldichte einer diskreten Zufallsvariablen M. Die Verteilungsfunktion

$$F(x) = \sum_{n=1}^{\infty} p_n F_n(x) \tag{3.3}$$

heißt **Mischung von Verteilungsfunktionen** F_n mit Gewichten p_n.

Übungsaufgabe 3.3.1 Zeigen Sie, dass F aus (3.3) eine gültige Verteilungsfunktion ist.

Bemerkung 3.3.4 a) Falls $p_n = 0$ für $n > N$, $N \in \mathbb{N}$, spricht man von einer **endlichen Mischung**

$$F(x) \sum_{n=1}^{N} p_n F_n(x)$$

b) Die Zufallsvariable X mit Verteilungsfunktion F aus (3.3) kann wie folgt simuliert werden:

1. Simuliere die Zufallsvariable M. Sei $M = n$ ihre Realisierung.
2. Simuliere die Zufallsvariable X mit Verteilungsfunktion F_n.

c) Generell können Mischungen einer parametrischen Familie $\{F_\mu\}$ von Verteilungen bzgl. des Parameters $\mu \in \mathbb{R}$ definiert werden, wenn μ als Realisierung einer Zufallsvariablen M mit der Verteilungsfunktion Φ_M aufgefasst wird:

$$F(x) = \int_{\mathbb{R}} F_\mu(x) d\Phi_M(\mu), \quad x \in \mathbb{R}.$$

Dabei wird die Borel-Messbarkeit von $F_\mu(x)$ bzgl. μ für alle $x \in \mathbb{R}$ vorausgesetzt.

3.4 Verteilungen von Zufallsvektoren

In der Definition 3.1.1, 2) wurden Zufallsvektoren bereits eingeführt. Sei $(\Omega, \mathcal{F}, P)$ ein beliebiger Wahrscheinlichkeitsraum und $X : \Omega \to \mathbb{R}^n$ ein n–dimensionaler Zufallsvektor, wobei wir seine Koordinaten als $(X_1, \ldots, X_n)$ bezeichnen. Dann folgt aus Definition 3.1.1, 2), dass $X_i, \quad i = 1, \ldots, n$ Zufallsvariablen sind. Umgekehrt kann man einen beliebigen Zufallsvektor X definieren, indem man seine Koordinaten $X_1 \ldots X_n$ als Zufallsvariablen auf demselben Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, P)$ einführt (Übungsaufgabe).

Definition 3.4.1 Sei $X = (X_1, \ldots, X_n)$ ein Zufallsvektor auf $(\Omega, \mathcal{F}, P)$.

1. Die *Verteilung* von X ist die Mengenfunktion $P_X : \mathcal{B}_{\mathbb{R}^n} \to [0, 1]$ mit $P_X(B) = P(X \in B) = P(\{\omega \in \Omega : X(\omega) \in B\}), \quad B \in \mathcal{B}_{\mathbb{R}^n}$.

2. Die *Verteilungsfunktion* $F_X : \mathbb{R}^n \to [0, 1]$ von X ist gegeben durch $F_X(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_N \leq x_n) \quad x_1, \ldots, x_n \in \mathbb{R}$. Sie heißt manchmal auch die *gemeinsame* oder die *multivariate Verteilungsfunktion* von X, um sie von folgenden *marginalen Verteilungsfunktionen* zu unterscheiden.

3. Sei $\{i_1, \ldots, i_k\}$ ein Teilvektor von $\{1, \ldots, n\}$. Die multivariate Verteilungsfunktion $F_{i_1,\ldots,i_k}$ des Zufallsvektors $(X_{i_1}, \ldots, X_{i_k})$ heißt *marginale Verteilungsfunktion* von X. Insbesondere für $k = 1$ und $i_1 = i$ spricht man von den so genannten *Randverteilungen*:

$$F_{X_i}(x) = P(X_i \leq x), \quad i = 1, \ldots, n\,.$$

Satz 3.4.1 *(Eigenschaften multivariater Verteilungsfunktionen):*
Sei $F_X : \mathbb{R}^n \to [0, 1]$ die Verteilungsfunktion eines Zufallsvektors $X = (X_1, \ldots, X_n)$. Dann gelten folgende Eigenschaften:

1. *Asymptotik:*

$$\lim_{x_i \to -\infty} F_X(x_1, \ldots, x_n) = 0, \quad \forall i = 1, \ldots, n \quad \forall x_1, \ldots, x_n \in \mathbb{R}\,,$$

$$\lim_{x_1, \ldots, x_n \to +\infty} F_X(x_1, \ldots, x_n) = 1\,,$$

$$\lim_{x_j \to +\infty \forall j \notin \{i_1, \ldots, i_k\}} F_X(x_1, \ldots, x_n) = F_{(X_{i_1}, \ldots, X_{i_k})}(x_{i_1}, \ldots, x_{i_k})\,,$$

wobei $F_{(X_{i_1}, \ldots, X_{i_k})}(x_{i_1}, \ldots, x_{i_k})$ die Verteilungsfunktion der marginalen Verteilung von

$$(X_{i_1} \ldots X_{i_k}) \text{ mit } \{i_1, \ldots, i_k\} \subset \{1, \ldots, n\} \text{ ist.}$$

Insbesondere gilt

$$\lim_{x_j \to +\infty,\, j \neq i} F_X(x_1, \ldots, x_n) = F_{X_i}(x_i), \quad \forall i = 1, \ldots, n$$

(Randverteilungsfunktion).

2. *Monotonie:* $\forall (x_1, \ldots, x_n) \in \mathbb{R}^n \quad \forall h_1, \ldots, h_n \geq 0$

$$F_X(x_1 + h_1, \ldots, x_n + h_n) \geq F_X(x_1, \ldots, x_n).$$

3. *Rechtsseitige Stetigkeit:*

$$F_X(x_1, \ldots, x_n) = \lim_{y_i \to x_i + 0,\, i = 1, \ldots, n} F_X(y_1, \ldots, y_n) \quad \forall (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

Beweis Analog zum Satz 3.2.1. □

Definition 3.4.2 Die Verteilung eines Zufallsvektors $X = (X_1, \ldots, X_n)$ heißt

1. *diskret*, falls eine höchstens abzählbare Menge $C \subset \mathbb{R}^n$ existiert, für die $P(X \in C) = 1$ gilt. Die Familie von Wahrscheinlichkeiten

$$\{P(X = x), x \in C\}$$

heißt dann *Wahrscheinlichkeitsfunktion* oder *Zähldichte* von X.

2. *absolut stetig*, falls eine Funktion $f_X : \mathbb{R}^n \to [0, 1]$ existiert, die Lebesgue–integrierbar auf $\mathbb{R}^n$ ist und für die gilt

$$F_X(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_X(y_1, \ldots, y_n)\, dy_n \ldots dy_1,$$

$\forall (x_1, \ldots, x_n) \in \mathbb{R}^n$. f_X heißt *Dichte* der gemeinsamen Verteilung von X.

Lemma 3.4.1 Sei $X = (X_1, \ldots, X_n)$ ein diskreter (bzw. absolut stetiger) Zufallsvektor mit Zähldichte $P(X = x)$ (bzw. Dichte $f_X(x)$). Dann gilt:

1. Die Verteilung P_X von X ist gegeben durch

$$P_X(B) = \sum_{x \in B} P(X = x) \text{ bzw. } P_X(B) = \int_B f_X(x)\, dx, \quad B \in \mathcal{B}_{\mathbb{R}^n}.$$

2. Die Koordinaten X_i, $i = 1, \ldots, n$ sind ebenfalls diskrete bzw. absolut stetige Zufallsvariablen mit der Randzähldichte

$$P(X_i = x) = \sum_{(y_1,\ldots,y_{i-1},x,y_{i+1},\ldots,y_n) \in C} P(X_1 = y_1, \ldots, X_{i-1} = y_{i-1}, X_i = x, X_{i+1} = y_{i+1}, \ldots, X_n = y_n)$$

bzw. Randdichte

$$f_{X_i}(x) = \int_{\mathbb{R}^{n-1}} f_X(y_1, \ldots, y_{i-1}, x, y_{i+1}, \ldots, y_n)\, dy_1 \ldots dy_{i-1}\, dy_{i+1} \ldots dy_n$$

$\forall x \in \mathbb{R}$.

Beweis

1. Folgt aus dem eindeutigen Zusammenhang zwischen einer Verteilung und ihrer Verteilungsfunktion.

2. Die Aussage für diskrete Zufallsvektoren ist trivial. Sei nun $X = (X_1, \ldots, X_n)$ absolut stetig. Dann folgt aus Satz 3.4.1

$$\begin{aligned} F_{X_i}(x) &= \lim_{y_j \to +\infty, j \neq i} F_X(x_1 \ldots x_{i-1}, x, x_{i+1}, \ldots, x_n) \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{-\infty}^{x} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(y_1, \ldots, y_n)\, dy_n \ldots dy_1 \\ &\underbrace{=}_{\text{S. v. Fubini}} \int_{-\infty}^{x} \Big(\underbrace{\int_{\mathbb{R}^{n-1}} f_X(y_1, \ldots, y_n)\, dy_1 \ldots dy_{i-1}\, dy_{i+1} \ldots dy_n}_{f_{X_i}(y_i)} \Big)\, dy_i \end{aligned}$$

Somit ist X_i absolut stetig verteilt mit Dichte f_{X_i}.

□

Beispiel 3.4.1 *Verschiedene Zufallsvektoren:*

1. *Polynomiale Verteilung:*
$X = (X_1, \ldots, X_k) \sim \text{Polynom}(n, p_1, \ldots, p_k)$, $n \in \mathbb{N}$, $p_i \in [0, 1]$, $i = 1, \ldots, k$, $\sum_{i=1}^n p_i = 1$, falls X diskret verteilt ist mit Zähldichte

$$P(X = x) = P(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k}$$

$\forall x = (x_1, \ldots, x_k)$ mit $x_i \in \mathbb{N} \cup \{0\}$ *und* $\sum_{i=1}^{k} x_i = n$. Die polynomiale Verteilung ist das k–dimensionale Analogon der Binomialverteilung. So sind die Randverteilungen von $X_i \sim Bin(n, p_i), \quad i = 1, \ldots, k$. (Bitte prüfen Sie dies als Übungsaufgabe!). Es gilt $P(\sum_{i=1}^{k} X_i = n) = 1$.

Interpretation:
Es werden n Versuche durchgeführt. In jedem Versuch kann eines aus insgesamt k Merkmalen auftreten. Sei p_i die Wahrscheinlichkeit des Auftretens von Merkmal i in einem Versuch. Sei

$$X_i = \#\{\text{Auftretens von Merkmal } i \text{ in } n \text{ Versuchen}\}, \quad i = 1, \ldots, k\,.$$

Dann ist $X = (X_1, \ldots, X_k) \sim \text{Polynom}(n, p_1, \ldots, p_k)$.

2. *Gleichverteilung:*
$X \sim \mathcal{U}(A)$, wobei $A \subset \mathbb{R}^n$ eine beschränkte Borel–Teilmenge von $\mathbb{R}^n$ ist, falls $X = (X_1, \ldots, X_n)$ absolut stetig verteilt mit der Dichte

$$f_X(x_1, \ldots, x_n) = \begin{cases} \frac{1}{|A|}, & (x_1, \ldots, x_n) \in A\,, \\ 0, & \text{sonst}, \end{cases}$$

ist (vgl. Abb. 3.13). Im Spezialfall $A = \prod_{i=1}^{n} [a_i, b_i]$ (Parallelepiped)

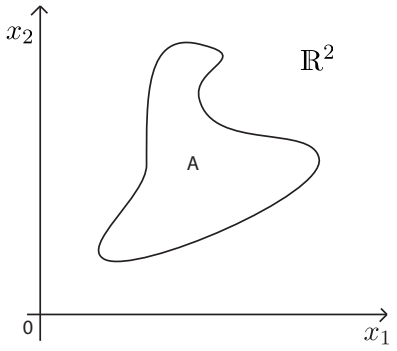


Abbildung 3.13: Wertebereich A einer zweidimensionalen Gleichverteilung

sind alle Randverteilungen von X_i ebenso Gleichverteilungen:

$$X_i \sim U[a_i, b_i], \quad i = 1, \ldots, n.$$

Interpretation:
$X = (X_1, \ldots, X_n)$ sind Koordinaten eines zufälligen Punktes, der gleichwahrscheinlich auf A geworfen wird. Dies ist die geometrische Wahrscheinlichkeit, denn $P(X \in B) = \int_B f_X(y)\, dy = \frac{|B|}{|A|}$ für $B \in \mathcal{B}_{\mathbb{R}^n} \cap A$.

3. *Multivariate Normalverteilung:*
   $X = (X_1, \ldots, X_n) \sim N(\mu, K)$, $\mu \in \mathbb{R}^n$, K eine positiv definite $(n \times n)$–Matrix, falls X absolut stetig verteilt mit Dichte

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n \cdot \det K}} \exp\{-\frac{1}{2}(X-\mu)^T K^{-1}(X-\mu)\}, \quad x \in \mathbb{R}^n$$

   ist.

   Spezialfall zweidimensionale Normalverteilung:

   Falls $n = 2$ und

$$\mu = (\mu_1, \mu_2),\ K = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

   dann gilt $\det K = \sigma_1^2\sigma_2^2(1-\rho^2)$ und

$$f_X(x_1, x_2) = \frac{1}{\sqrt{1-\rho^2} \cdot 2\pi\sigma_1\sigma_2} \times$$
$$\times \exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right)\right\},$$

   $(x_1, x_2) \in \mathbb{R}^2$, weil

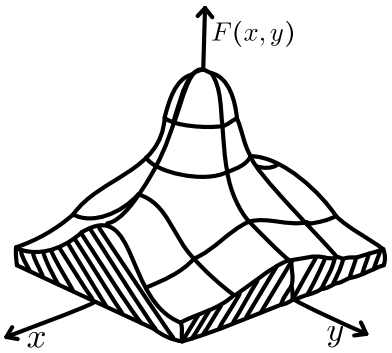


Abbildung 3.14: Grafik der Dichte einer zweidimensionalen Normalverteilung

$$K^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix} \frac{1}{\sigma_1^2} & -\frac{\rho}{\sigma_1\sigma_2} \\ -\frac{\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{pmatrix}, \quad \text{(vgl. Abb. 3.14).}$$

Übungsaufgabe 3.4.1 Zeigen Sie, dass $X_i \sim N(\mu_i, \sigma_i^2), \quad i = 1, 2$. Diese Eigenschaft der Randverteilungen gilt für alle $n \geq 2$. Somit ist die multivariate Normalverteilung ein mehrdimensionales Analogon der eindimensionalen $N(\mu, \sigma^2)$–Verteilung.

Interpretation:
Man feuert eine Kanone auf das Ziel mit Koordinaten (μ_1, μ_2). Dann sind $X = (X_1, X_2)$ die Koordinaten des Treffers. Durch die Streuung gilt, dass $(X_1, X_2) = (\mu_1, \mu_2)$ nur im Mittel eintritt. Die Varianzen σ_1^2 und σ_2^2 sind Maße für die Genauigkeit des Feuers.

3.5 Stochastische Unabhängigkeit

3.5.1 Unabhängige Zufallsvariablen

Definition 3.5.1

1. Seien $X_1, \ldots, X_n$ Zufallsvariablen definiert auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, P)$. Sie heißen *stochastisch unabhängig*, falls

$$F_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdot \ldots \cdot F_{X_n}(x_n), \quad x_1, \ldots, x_n \in \mathbb{R}$$

   oder äquivalent dazu,

$$P(X_1 \leq x_1, \cdots, X_n \leq x_n) = P(X_1 \leq x_1) \cdot \ldots \cdot P(X_n \leq x_n)\,.$$

2. Sei $\{X_n\}_{n\in\mathbb{N}}$ eine Folge von Zufallsvariablen definiert auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, P)$. Diese Folge besteht aus *stochastisch unabhängigen Zufallsvariablen*, falls $\forall k \in \mathbb{N}\, \forall i_1 < i_2 < \cdots < i_k\, X_{i_1}, X_{i_2}, \ldots, X_{i_k}$ stochastisch unabhängige Zufallsvariablen (im Sinne der Definition 1) sind.

Lemma 3.5.1 Die Zufallsvariablen $X_1, \ldots, X_n$ sind genau dann stochastisch unabhängig, wenn für alle $B_1, \ldots, B_n \in \mathcal{B}_\mathbb{R}$ gilt

$$P(X_1 \in B_1, \ldots, X_n \in B_n) = P(X_1 \in B_1) \cdot \ldots \cdot P(X_n \in B_n) \tag{3.4}$$

Satz 3.5.1 *(Charakterisierung der stochastischen Unabhängigkeit von Zufallsvariablen)*

1. Sei $(X_1, \ldots, X_n)$ ein diskret verteilter Zufallsvektor mit dem Wertebereich C. Seine Koordinaten $X_1, \ldots, X_n$ sind genau dann stochastisch unabhängig, wenn

$$P(X_1 = x_1, \ldots, X_n = x_n) \;=\; \prod_{i=1}^{n} P(X_i = x_i)$$

   $\forall (x_1, \ldots, x_n) \in C$.

2. Sei $(X_1, \ldots, X_n)$ ein absolut stetiger Zufallsvektor mit der Wahrscheinlichkeitsdichte $f_{(X_1 \ldots X_n)}$ und Randdichten f_{X_i}. Es gilt, dass die Koordinaten $X_1, \ldots, X_n$ genau dann stochastisch unabhängig sind, wenn

$$f_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

für fast alle $(x_1, \ldots, x_n) \in \mathbb{R}^n$.

Beispiel 3.5.1 1. *Multivariate Normalverteilung:*
Mit Hilfe des Satzes 3.5.1 kann gezeigt werden, dass die Komponenten $X_1, \ldots, X_n$ von

$$X = (X_1, \ldots, X_n) \sim N(\mu, K)$$

genau dann unabhängig sind, wenn $k_{ij} = 0$, $i \neq j$, wobei $K = (k_{ij})_{i,j=1}^n$. Insbesondere gilt im zweidimensionalen Fall (vgl. Bsp. 3 Seite 45), dass X_1 und X_2 unabhängig sind, falls $\rho = 0$.

Übungsaufgabe 3.5.1 Zeigen Sie es!

2. *Multivariate Gleichverteilung:*
Die Komponenten des Vektors $X = (X_1, \ldots, X_n) \sim \mathcal{U}(A)$ sind genau dann unabhängig, falls $A = \prod_{i=1}^n [a_i, b_i]$ ist. In der Tat gilt dann

$$f_X(x_1, \ldots, x_n) = \begin{cases} \frac{1}{|A|} = \frac{1}{\prod_{i=1}^n (b_i - a_i)} = \prod_{i=1}^{n} \frac{1}{b_i - a_i} = \prod_{i=1}^{n} f_{X_i}(x_i), & x \in A \\ 0, & \text{sonst} \end{cases}$$

mit $x = (x_1, \ldots, x_n)$, wobei

$$f_X(x_i) = \begin{cases} 0, & \text{falls } x_i \notin [a_i, b_i] \\ \frac{1}{b_i - a_i}, & \text{sonst}\,. \end{cases}$$

Implizit haben wir an dieser Stelle benutzt, dass

$$X_i \sim \mathcal{U}[a_i, b_i],\ i = 1, \ldots, n\ .$$

Herleitung:

$$\int_{\mathbb{R}^{n-1}} f_X(x)\, dx_n \ldots dx_{i+1}\, dx_{i-1} \ldots dx_1 = \frac{1}{b_i - a_i},\ x_i \in [a_i, b_i].$$

Übungsaufgabe 3.5.2
Zeigen Sie die Notwendigkeit der Bedingung $A = \prod_{i=1}^n [a_i, b_i]$!

3. Es gibt Beispiele von Zufallsvariablen X_1 und X_2, die stochastisch abhängig von einander sind, so dass X_1^2 stochastisch unabhängig von X_2^2 ist.
Unterschied: Kausale bzw. stochastische Abhängigkeit!

Definition 3.5.2 Eine Funktion $\varphi : \mathbb{R}^n \to \mathbb{R}^m$, $n, m \in \mathbb{N}$ heißt *Borelsche Funktion*, falls sie $\mathcal{B}_{\mathbb{R}^n}$–messbar ist, d.h. $\forall B \in \mathcal{B}_{\mathbb{R}^m}$ ist $\varphi^{-1}(B) \in \mathcal{B}_{\mathbb{R}^n}$.

Bemerkung 3.5.1

1. Sei $X = (X_1, \ldots, X_n)^T$ ein Zufallsvektor, und $\phi : \mathbb{R}^n \to \mathbb{R}^m$ sei Borelsch. Dann ist $Y = (Y_1, \ldots, Y_m)^T = \phi(X)$ ebenfalls ein Zufallsvektor.

2. Seien X_1, X_2 Zufallsvariablen auf $(\Omega, \mathcal{F}, P)$ und $\varphi_1, \varphi_2 : \mathbb{R} \to \mathbb{R}$ Borelsche Funktionen. Falls X_1 und X_2 stochastisch unabhängig sind, dann sind auch $\varphi_1(X_1)$ und $\varphi_2(X_2)$ stochastisch unabhängig.

3.6 Funktionen von Zufallsvektoren

Lemma 3.6.1 Jede stetige Funktion $\varphi : \mathbb{R}^n \to \mathbb{R}^m$ ist Borel–messbar. Insbesondere sind Polynome $\varphi : \mathbb{R}^n \to \mathbb{R}$ der Form

$$\varphi(x_1, \ldots, x_n) = \sum_{i=0}^{k} a_i x_1^{m_{1_i}} \ldots x_n^{m_{n_i}},$$

$k \in \mathbb{N}$, $a_0, a_1, \ldots, a_n \in \mathbb{R}$, $m_{1_i}, \ldots, m_{n_i} \in \mathbb{N} \cup \{0\}$, $i = 1, \ldots, k$ Borel–messbar.

Satz 3.6.1 (*Transformationssatz*)
Sei X eine Zufallsvariable auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, P)$.

1. Falls die Abbildung $\varphi : \mathbb{R} \to \mathbb{R}$ stetig und streng monoton wachsend ist, dann gilt $F_{\varphi(X)}(x) = F_X(\varphi^{-1}(x)) \quad \forall x \in \mathbb{R}$, wobei φ^{-1} die Umkehrfunktion von φ ist. Falls $\varphi : \mathbb{R} \to \mathbb{R}$ stetig und streng monoton fallend ist, dann gilt $F_{\varphi(X)}(x) = 1 - F_X(\varphi^{-1}(x)) + P(X = \varphi^{-1}(x))$, $x \in \mathbb{R}$.

2. Falls X absolut stetig mit Dichte f_X ist und $C \subset \mathbb{R}$ eine offene Menge mit $P(X \in C) = 1$ ist, dann ist $\varphi(X)$ absolut stetig mit Dichte $f_{\varphi(X)}(y) = f_X(\varphi^{-1}(y)) \cdot |(\varphi^{-1})'(y)|, \quad y \in \varphi(C)$, falls φ eine auf C stetig differenzierbare Funktion mit $\varphi'(x) \neq 0, \ x \in C$ ist.

Beweis

1. Falls φ monoton wachsend ist, gilt für $x \in \mathbb{R}$, dass
   $F_{\varphi(X)}(x) = P(\varphi(X) \leq x) = P(X \leq \varphi^{-1}(x)) = F_X(\varphi^{-1}(x))$.
   Für φ monoton fallend gilt für $x \in \mathbb{R}$

$$\begin{aligned} F_{\varphi(X)}(x) &= P(\varphi(X) \leq x) = P(X \geq \varphi^{-1}(x)) \\ &= 1 - P(X < \varphi^{-1}(x)) = 1 - F_X(\varphi^{-1}(x)) + P(X = \varphi^{-1}(x)). \end{aligned}$$

2. Nehmen wir o.B.d.A. an, dass $C = \mathbb{R}$ und $\varphi'(x) > 0 \, \forall x \in \mathbb{R}$.
   Für $\varphi'(x) < 0$ verläuft der Beweis analog. Aus Punkt 1) folgt

$$\begin{aligned} F_{\varphi(X)}(x) &= F_X(\varphi^{-1}(x)) \\ &= \int_{-\infty}^{\varphi^{-1}(x)} f_X(y)\, dy \\ &\underbrace{=}_{\varphi^{-1}(t)=y} \int_{-\infty}^{x} f_X(\varphi^{-1}(t)) \cdot |(\varphi^{-1})'(t)|\, dt, \quad x \in \mathbb{R}\,. \end{aligned}$$

   Hieraus folgt $f_{\varphi(X)}(t) = f_X(\varphi^{-1}(t)) \cdot |(\varphi^{-1})'(t)|, \quad t \in \mathbb{R}\,.$

□

Satz 3.6.2 *(Lineare Transformation)*
Sei $X : \Omega \to \mathbb{R}$ eine Zufallsvariable und $a, b \in \mathbb{R}$, $a \neq 0$. Dann gilt Folgendes:

1. Die Verteilungsfunktion der Zufallsvariable $aX + b$ ist gegeben durch

$$F_{aX+b}(x) = \begin{cases} F_X\left(\frac{x-b}{a}\right), & a > 0 \\ 1 - F_X\left(\frac{x-b}{a}\right) + P\left(X = \frac{x-b}{a}\right), & a < 0\,. \end{cases}$$

2. Falls X absolut stetig mit Dichte f_X ist, dann ist $aX + b$ ebenfalls absolut stetig mit Dichte

$$f_{aX+b}(x) = \frac{1}{|a|} f_X\left(\frac{x-b}{a}\right).$$

Beweis 1. Der Fall $a > 0$ $(a < 0)$ folgt aus dem Satz 3.6.1, 1), weil $\varphi(x) = aX + b$ stetig und monoton wachsend bzw. fallend ist.

2. Folgt aus dem Satz 3.6.1, 2), weil $\varphi(x) = aX + b$ stetig differenzierbar auf $C = \mathbb{R}$ (offen) mit $\varphi'(x) = a \neq 0$ ist.

□

Satz 3.6.3 *(Quadrierung)*
Sei X eine Zufallsvariable auf $(\Omega, \mathcal{F}, P)$.

1. Die Verteilungsfunktion von X^2 ist gegeben durch

$$F_{X^2}(x) = \begin{cases} F_X(\sqrt{x}) - F_X(-\sqrt{x}) + P(X = -\sqrt{x}), & \text{falls } x \geq 0 \\ 0, & \text{sonst}\,. \end{cases}$$

2. Falls X absolut stetig mit Dichte f_X ist, dann ist auch X^2 absolut stetig mit Dichte

$$f_{X^2}(x) = \begin{cases} \frac{1}{2\sqrt{x}} \left(f_X(\sqrt{x}) + f_X(-\sqrt{x})\right), & x > 0 \\ 0, & \text{sonst}\,. \end{cases}$$

Beweis 1. Für $x < 0$ gilt $F_{X^2}(x) = P(X^2 \leq x) = 0$.
Für $x \geq 0$ gilt

$$\begin{aligned} F_{X^2}(x) &= P(X^2 \leq x) = P(|X| \leq \sqrt{x}) \\ &= P(-\sqrt{x} \leq X \leq \sqrt{x}) = P(X \leq \sqrt{x}) - P(X < -\sqrt{x}) \\ &= F_X(\sqrt{x}) - F_X(-\sqrt{x}) + P(X = -\sqrt{x})\,. \end{aligned}$$

2. Wegen 1) gilt $F_{X^2}(x) = F_X(\sqrt{x}) - F_X(-\sqrt{x})$, da im absolut stetigen Fall $P(X = -\sqrt{x}) = 0 \quad \forall x \geq 0$. Daher gilt

$$\begin{aligned} F_{X^2}(x) &= \int_{-\sqrt{x}}^{\sqrt{x}} f_X(y)\,dy = \\ &= \int_0^{\sqrt{x}} f_X(y)\,dy + \int_{-\sqrt{x}}^0 f_X(y)\,dy \\ &\underbrace{=}_{y=\sqrt{t} \text{ oder } y=-\sqrt{t}} \int_0^x \frac{1}{2\sqrt{t}} f_X(\sqrt{t})\,dt + \int_0^x \frac{1}{2\sqrt{t}} f_X(-\sqrt{t})\,dt \\ &= \int_0^x \frac{1}{2\sqrt{t}} \left(f_X(\sqrt{t}) + f_X(-\sqrt{t})\right)\,dt\,, \forall x \geq 0\,. \end{aligned}$$

Daher gilt die Aussage 2) des Satzes.

□

Beispiel 3.6.1

1. Falls $X \sim N(0,1)$, dann ist $Y = \mu + \sigma X \sim N(\mu, \sigma^2)$.
2. Falls $X \sim N(\mu, \sigma^2)$, dann heißt die Zufallsvariable $Y = e^X$ *lognormalverteilt mit Parametern μ und σ^2*. Diese Verteilung wird sehr oft in ökonometrischen Anwendungen benutzt.

   Zeigen Sie, dass die Dichte von Y durch

$$f_Y(x) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

   gegeben ist.
3. Falls $X \sim N(0,1)$, dann heißt $Y = X^2$ χ_1^2*–verteilt (Chi-Quadrat-Verteilung) mit einem Freiheitsgrad.*

   Zeigen Sie, dass die Dichte von Y durch

$$f_Y(x) = \begin{cases} \frac{1}{\sqrt{2\pi x}} e^{-\frac{x}{2}}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

gegeben ist.

Die χ^2-Verteilung wird in der Statistik sehr oft als die sogenannte *Prüfverteilung* bei der Konstruktion von statistischen Tests und Konfidenzintervallen verwendet.

Satz 3.6.4 *(Summe von unabhängigen Zufallsvariablen)*
Sei $X = (X_1, X_2)$ ein absolut stetiger Zufallsvektor mit Dichte f_X. Dann gilt Folgendes:

1. Die Zufallsvariable $Y = X_1 + X_2$ ist absolut stetig mit Dichte

$$f_Y(x) = \int_{\mathbb{R}} f_X(y, x-y)\, dy, \quad \forall x \in \mathbb{R}\,. \tag{3.5}$$

2. Falls X_1 und X_2 unabhängig sind, dann heißt der Spezialfall

$$f_Y(x) = \int_{\mathbb{R}} f_{X_1}(y) f_{X_2}(x-y)\, dy, \quad x \in \mathbb{R}$$

von (3.5) *Faltungsformel*.

Beweis
2) ergibt sich aus 1) für $f_X(x,y) = f_{X_1}(x) \cdot f_{X_2}(y) \quad \forall x, y \in \mathbb{R}$.

Beweisen wir also 1):

$$\begin{aligned}
P(Y \le t) &= P(X_1 + X_2 \le t) = \int_{(x,y)\in\mathbb{R}^2 :\, x+y\le t} f_X(x,y)\, dx\, dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{t-x} f_X(x,y)\, dy\, dx \\
&\underbrace{=}_{y \mapsto z = x+y} \int_{-\infty}^{\infty} \int_{-\infty}^{t} f_X(x, z-x)\, dz\, dx \\
&\underbrace{=}_{\text{Fubini}} \int_{-\infty}^{t} \underbrace{\int_{\mathbb{R}} f_X(x, z-x)\, dx}_{= f_Y(z)}\, dz, t \in \mathbb{R}.
\end{aligned}$$

Somit ist $f_Y(z) = \int_{\mathbb{R}} f_X(x, z-x)\, dx$ die Dichte von $Y = X_1 + X_2$. □

Folgerung 3.6.1 *(Faltungsstabilität der Normalverteilung):*
Falls die Zufallsvariablen $X_1, \ldots, X_n$ unabhängig mit

$$X_i \sim N(\mu_i, \sigma_i^2) \quad i = 1, \ldots, n$$

sind, dann gilt

$$X_1 + \cdots + X_n \sim N(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2)\,.$$

Beweis Induktion bzgl. n. Den Fall $n = 2$ sollten Sie als Übungsaufgabe lösen. Der Rest des Beweises ist trivial. □

Satz 3.6.5 *(Produkt und Quotient von Zufallsvariablen):*
Sei $X = (X_1, X_2)$ ein absolut stetiger Zufallsvektor mit Dichte f_X. Dann gilt Folgendes:

1. Die Zufallsvariable $Y = X_1 \cdot X_2$ und $Z = \frac{X_1}{X_2}$ sind absolut stetig verteilt mit Dichten
$$f_Y(x) = \int_{\mathbb{R}} \frac{1}{|t|} f_X\left(x/t, t\right)\, dt,$$
bzw.
$$f_Z(x) = \int_{\mathbb{R}} |t| f_X(x \cdot t, t)\, dt, x \in \mathbb{R}.$$

2. Falls X_1 und X_2 unabhängig sind, dann gilt der Spezialfall der obigen Formeln
$$f_Y(x) = \int_{\mathbb{R}} \frac{1}{|t|} f_{X_1}\left(x/t\right) f_{X_2}\left(t\right)\, dt,$$
bzw.
$$f_Z(x) = \int_{\mathbb{R}} |t| f_{X_1}(x \cdot t) f_{X_2}(t)\, dt, x \in \mathbb{R}.$$

Beweis Analog zu dem Beweis des Satzes 3.6.4. □

Beispiel 3.6.2 Zeigen Sie, dass $X_1/X_2 \sim$ Cauchy(0,1), falls $X_1, X_2 \sim N(0,1)$ und unabhängig sind:

$$f_{X_1/X_2}(x) = \frac{1}{\pi(x^2+1)}, \quad x \in \mathbb{R}\,.$$

Bemerkung 3.6.1 Da X_1 und X_2 absolut stetig verteilt sind, tritt das Ereignis $\{X_2 = 0\}$ mit Wahrscheinlichkeit Null ein. Daher ist $X_1(\omega)/X_2(\omega)$ wohl definiert für fast alle $\omega \in \Omega$. Für $\omega \in \Omega : X_2(\omega) = 0$ kann $X_1(\omega)/X_2(\omega)$ z.B. als 1 definiert werden. Dies ändert den Ausdruck der Dichte von X_1/X_2 nicht.

Kapitel 4

Momente von Zufallsvariablen

Weitere wichtige Charakteristiken von Zufallsvariablen sind ihre so genannten *Momente*, darunter der Erwartungswert und die Varianz. Zusätzlich wird in diesem Kapitel die Kovarianz von zwei Zufallsvariablen als Maß ihrer Abhängigkeit diskutiert. Um diese Charakteristiken einführen zu können, brauchen wir die Definition des Lebesgue–Integrals auf beliebigen messbaren Räumen.

Beispiel 4.0.1 Sei X eine diskrete Zufallsvariable mit dem endlichen Wertebereich $C = \{x_1, \ldots, x_n\}$ und Zähldichte $\{p_i\}_{i=1}^n$, wobei $p_i = P(X = x_i)$, $i = 1, \ldots, n$. Wie soll der Mittelwert von X berechnet werden? Aus der Antike sind drei Ansätze zur Berechnung des Mittels von n Zahlen bekannt:

- das arithmetische Mittel: $\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$
- das geometrische Mittel: $\bar{g}_n = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$
- das harmonische Mittel: $\bar{h}_n = \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{x_i} \right)^{-1}$

Um $\bar{g}_n$ und $\bar{h}_n$ berechnen zu können, ist die Bedingung $x_i > 0$ bzw. $x_i \neq 0$ $i = 1, \ldots, n$ eine wichtige Voraussetzung. Wir wollen jedoch diese Mittel für beliebige Wertebereiche einführen. Somit fallen diese zwei Möglichkeiten schon weg. Beim arithmetischen Mittel werden beliebige x_i zugelassen, jedoch alle gleich gewichtet, unabhängig davon, ob der Wert x_{i_0} wahrscheinlicher als alle anderen Werte ist und somit häufiger in den Experimenten vorkommt.

Deshalb ist es naheliegend, das gewichtete Mittel $\sum_{i=1}^n x_i \omega_i$ zu betrachten, $\forall i \ \omega_i \geq 0$, $\sum_{i=1}^n \omega_i = 1$, wobei das Gewicht ω_i die relative Häufigkeit des Vorkommens des Wertes x_i in den Experimenten ausdrückt. Somit ist es natürlich, $\omega_i = p_i$ zu setzen, $i = 1, \ldots, n$, und schreiben $\mathrm{E}X = \sum_{i=1}^n x_i p_i$.

Dieses Mittel wird "Erwartungswert der Zufallsvariable X" genannt. Der Buchstabe "$\mathbb{E}$" kommt aus dem Englischen: "Expectation". Für die Gleichverteilung auf $\{x_1, \ldots, x_n\}$, d.h. $p_i = \frac{1}{n}$, stimmt $\mathbb{E}X$ mit dem arithmetischen Mittel $\bar{x}_n$ überein. Wie wir bald sehen werden, kann

$$\mathbb{E}X = \sum_{i=1}^{n} x_i P(X = x_i)$$

geschrieben werden.

4.1 Erwartungswert

Somit können wir folgende Definition angeben:

Definition 4.1.1

1. Sei X eine diskret verteilte Zufallsvariable mit Wertebereich C und Zähldichte $P_X(x)$. Der Erwartungswert von X ist definiert als

$$\mathbb{E}X = \sum_{x \in C} x \; P_X(x),$$

   falls $\sum\limits_{x \in C} |x| \; P_X(x) < \infty$.

2. Sei X absolut stetig verteilt mit Dichte f_X. Der Erwartungswert von X ist definiert als

$$\mathrm{E}X = \int_{\mathbb{R}} x \; f_X(x) \; dx,$$

   falls $\underbrace{\int_{\mathbb{R}} |x| \; f_X(x) \; dx}_{\mathrm{E}|X|} < \infty$

Satz 4.1.1 *(Eigenschaften des Erwartungswertes)*
Seien X, Y integrierbare Zufallsvariablen auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, P)$. Dann gilt Folgendes:

1. Falls $X(\omega) = I_A(\omega)$ für ein $A \in \mathcal{F}$, dann gilt $\mathrm{E}X = P(A)$.
2. Falls $X(\omega) \geq 0,$ für fast alle $\omega \in \Omega$, dann ist $\mathrm{E}X \geq 0$.
3. *Additivität:* für beliebige $a, b \in \mathbb{R}$ gilt $\mathrm{E}(aX + bY) = a\mathrm{E}X + b\mathrm{E}Y$.
4. *Monotonie:* Falls $X \geq Y$ für fast alle $\omega \in \Omega$ (man sagt dazu *"fast sicher"* und schreibt *"f.s."*), dann gilt $\mathrm{E}X \geq \mathrm{E}Y$.
   Falls $0 \leq X \leq Y$ fast sicher und lediglich vorausgesetzt wird, dass Y integrierbar ist, dann ist auch X integrierbar.

5. $|\mathrm{E}X| \leq \mathrm{E}|X|$.
6. Falls X fast sicher auf Ω beschränkt ist, dann ist X integrierbar.
7. Falls X und Y unabhängig sind, dann gilt $\mathrm{E}(XY) = \mathrm{E}X \cdot \mathrm{E}Y$.
8. Falls $X \geq 0$ fast sicher und $\mathrm{E}X = 0$, dann gilt $X = 0$ fast sicher.

Bemerkung 4.1.1

1. Aus dem Satz 4.1.1, 3) und 7) folgt per Induktion, dass
   (a) Falls $X_1, \ldots, X_n$ integrierbare Zufallsvariablen sind und die Koeffizienten $a_1, \ldots, a_n \in \mathbb{R}$, dann ist $\sum_{i=1}^n a_i X_i$ eine integrierbare Zufallsvariable und es gilt
   $$\mathrm{E}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i \mathrm{E}X_i\,.$$
   (b) Falls $X_1, \ldots, X_n$ zusätzlich unabhängig sind und das Produkt $X_1 \cdot X_n$ integrierbar ist, d.h. $\mathrm{E}|X_1 \ldots X_n| < \infty$, dann gilt
   $$\mathrm{E}\left(\prod_{i=1}^n X_i\right) = \prod_{i=1}^n \mathrm{E}X_i\,.$$
2. Die Aussage 7) des Satzes 4.1.1 gilt nicht in umgekehrte Richtung: aus $\mathrm{E}(XY) = \mathrm{E}X \cdot \mathrm{E}Y$ folgt im Allgemeinen nicht die Unabhängigkeit von Zufallsvariablen X und Y. Als Illustration betrachten wir folgendes Beispiel:
   (a) Seien X_1, X_2 unabhängige Zufallsvariablen mit $\mathrm{E}X_1 = \mathrm{E}X_2 = 0$. Setzen wir $X = X_1$ und $Y = X_1 \cdot X_2$. X und Y sind abhängig und dennoch
   $$\mathrm{E}(XY) = \mathrm{E}(X_1^2 X_2) = \mathrm{E}X_1^2 \cdot \mathrm{E}X_2 = 0 = \mathrm{E}X \cdot \mathrm{E}Y = 0 \cdot \mathrm{E}Y = 0\,.$$
   (b) Falls der Zufallsvektor (X, Y) normalverteilt ist, dann sind X und Y unabhängig genau dann, wenn $\mathrm{E}(XY) = \mathrm{E}X \cdot \mathrm{E}Y$.

Folgerung 4.1.1

1. Falls X absolut stetig verteilt mit Dichte f_X ist, dann gilt
   $$\mathrm{E}g(X) = \int_{\mathbb{R}^n} g(x) f_X(x)\, dx\,.$$
2. Falls X diskret verteilt mit dem Wertebereich $C = \{x_1, x_2, \ldots\} \subset \mathbb{R}^n$ ist, dann gilt
   $$\mathrm{E}g(X) = \sum_i g(x_i) P(X = x_i) = \sum_{x \in C} g(x) P(X = x)\,.$$

Beispiele für die Berechnung des Erwartungswertes:

1. *Poisson–Verteilung:* Sei $X \sim \text{Poisson}(\lambda)$, $\lambda > 0$. Dann gilt

$$\begin{aligned} \mathrm{E}X &= \sum_{k=0}^{\infty} kP(X=k) = \sum_{k=0}^{\infty} ke^{-\lambda}\frac{\lambda^k}{k!} \\ &= e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}\cdot\lambda}{(k-1)!} \\ &\underset{k-1=n}{=} e^{-\lambda}\cdot\lambda\cdot\sum_{n=0}^{\infty}\frac{\lambda^n}{n!} = e^{-\lambda}\lambda e^{\lambda} = \lambda \Longrightarrow \mathrm{E}X = \lambda\,. \end{aligned}$$

2. *Normalverteilung:* Sei $X \sim N(\mu,\sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$. Zeigen wir, dass $\mathrm{E}X = \mu$.

$$\begin{aligned} \mathrm{E}X &= \int_{\mathbb{R}} x f_X(x)\,dx = \frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot\int_{-\infty}^{\infty} xe^{-\left(\frac{x-\mu}{\sigma}\right)^2\cdot\frac{1}{2}}\,dx \\ &\underset{y=\frac{x-\mu}{\sigma}}{=} \frac{1}{\sqrt{2\pi}}\cdot\int_{-\infty}^{\infty}(\sigma y+\mu)e^{-\frac{y^2}{2}}\,dy \\ &= \frac{\sigma}{\sqrt{2\pi}}\underbrace{\int_{\mathbb{R}} ye^{-\frac{y^2}{2}}\,dy}_{=0} + \frac{\mu}{\sqrt{2\pi}}\underbrace{\int_{\mathbb{R}} e^{-\frac{y^2}{2}}\,dy}_{=\sqrt{2\pi}} = \mu\,, \end{aligned}$$

weil

$$\int_{\mathbb{R}} ye^{-\frac{y^2}{2}}\,dy \underset{t=\frac{y^2}{2}}{=} \left(\int_{-\infty}^{0} + \int_{0}^{+\infty}\right)e^{-t}\,dt = -\left(\int_{0}^{+\infty} - \int_{0}^{+\infty} e^{-t}\right) = 0;$$

$$\begin{aligned} \left(\int_{\mathbb{R}} e^{-\frac{y^2}{2}}\,dy\right)^2 &= \int_{\mathbb{R}} e^{-\frac{-x^2}{2}}\,dx\cdot\int_{\mathbb{R}} e^{-\frac{y^2}{2}}\,dy \\ &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-\frac{x^2+y^2}{2}}\,dx\,dy \\ &\underset{(x,y)\mapsto\text{Polarkoord. }(r,\varphi)}{=} \int_0^{2\pi}\int_0^{+\infty} e^{-\frac{r^2}{2}}r\,dr\,d\varphi \\ &= 2\pi\cdot\int_0^{+\infty} e^{-\frac{r^2}{2}}\,d\left(\frac{r^2}{2}\right) \\ &\underset{\frac{r^2}{2}=t}{=} 2\pi(-1)\int_0^{+\infty} d\left(e^{-t}\right) = 2\pi \end{aligned}$$

$$\Longrightarrow \int_{\mathbb{R}} e^{-\frac{y^2}{2}}\,dy = \sqrt{2\pi} \quad \Longrightarrow \mathrm{E}X = \mu\,.$$

4.2 Varianz

Neben dem "Mittelwert" einer Zufallsvariablen, den der Erwartungswert repräsentiert, gibt es weitere Charakteristiken, die für die praktische Beschreibung der zufälligen Vorgänge in der Natur und Technik sehr wichtig sind. Die Varianz z.B. beschreibt die Streuung der Zufallsvariablen um ihren Mittelwert. Sie wird als mittlere quadratische Abweichung vom Erwartungswert eingeführt:

Definition 4.2.1 Sei X eine Zufallsvariable mit $\mathrm{E}X^2 < \infty$.

1. Die *Varianz* der Zufallsvariablen X wird als $\mathrm{Var}\ X = \mathrm{E}(X - \mathrm{E}X)^2$ definiert.
2. $\sqrt{\mathrm{Var}\ X}$ heißt *Standardabweichung* von X.
3. Seien X und Y zwei Zufallsvariablen mit $\mathrm{E}|XY| < \infty$. Die Größe $\mathrm{Cov}\ (X, Y) = \mathrm{E}(X - \mathrm{E}X)(Y - \mathrm{E}Y)$ heißt *Kovarianz* der Zufallsvariablen X und Y.
4. Falls $\mathrm{Cov}\ (X, Y) = 0$, dann heißen die Zufallsvariablen X und Y *unkorreliert*.

Satz 4.2.1 *(Eigenschaften der Varianz und der Kovarianz)*
Seien X, Y zwei Zufallsvariablen mit $\mathrm{E}(X^2) < \infty$, $\mathrm{E}(Y^2) < \infty$. Dann gelten folgende Eigenschaften:

1. $$\mathrm{Cov}\ (X, Y) = \mathrm{E}(XY) - \mathrm{E}X \cdot \mathrm{E}Y\,, \quad \mathrm{Var}\ X = \mathrm{E}(X^2) - (\mathrm{E}X)^2\,. \tag{4.1}$$
2. $$\mathrm{Cov}\ (aX + b, cY + d) = ac \cdot \mathrm{Cov}\ (X, Y), \quad \forall a, b, c, d \in \mathbb{R}\,, \tag{4.2}$$
$$\mathrm{Var}\ (aX + b) = a^2 \mathrm{Var}\ (X)\,, \quad \forall a, b \in \mathbb{R}\,. \tag{4.3}$$
3. $\mathrm{Var}\ X \geq 0$. Es gilt $\mathrm{Var}\ X = 0$ genau dann, wenn $X = \mathrm{E}X$ fast sicher.
4. $\mathrm{Var}\ (X + Y) = \mathrm{Var}\ X + \mathrm{Var}\ Y + 2\mathrm{Cov}\ (X, Y)$.
5. Falls X und Y unabhängig sind, dann sind sie unkorreliert, also gilt $\mathrm{Cov}\ (X, Y) = 0$.

Beweis 1. Beweisen wir die Formel $\mathrm{Cov}\ (X, Y) = \mathrm{E}(XY) - \mathrm{E}X \cdot \mathrm{E}Y$. Die Darstellung (4.1) für die Varianz ergibt sich aus dieser Formel für $X = Y$.

$$\begin{aligned}
\mathrm{Cov}\ (X, Y) &= \mathrm{E}(X - \mathrm{E}X)(Y - \mathrm{E}Y) \\
&= \mathrm{E}(XY - X\mathrm{E}Y - Y\mathrm{E}X + \mathrm{E}X \cdot \mathrm{E}Y) \\
&= \mathrm{E}(XY) - \mathrm{E}X \cdot \mathrm{E}Y - \mathrm{E}Y \cdot \mathrm{E}X + \mathrm{E}X \cdot \mathrm{E}Y \\
&= \mathrm{E}(XY) - \mathrm{E}X \cdot \mathrm{E}Y\,.
\end{aligned}$$

2. Beweisen wir die Formel (4.2). Die Formel (4.3) folgt aus (4.2) für $X = Y$, $a = c$ und $b = d$. Es gilt

$$\begin{aligned}\text{Cov}\,(aX+b, cY+d) &= \text{E}\,((aX+b-a\text{E}X-b)(cY+d-c\text{E}Y-d)) \\ &= \text{E}\,(ac(X-\text{E}X)(Y-\text{E}Y)) \\ &= ac\,\text{Cov}\,(X,Y)\,, \quad \forall\, a,b,c,d \in \mathbb{R}\,.\end{aligned}$$

3. Es gilt offensichtlich

$$\text{Var}\,X = \text{E}(X-\text{E}X)^2 \geq 0, \text{ da } (X-\text{E}X)^2 \geq 0 \quad \forall \omega \in \Omega\,.$$

Falls $X = \text{E}X$ fast sicher, dann gilt $(X-\text{E}X)^2 = 0$ fast sicher und somit $\text{E}(X-\text{E}X)^2 = 0 \Longrightarrow \text{Var}\,X = 0$.
Falls umgekehrt $\text{Var}\,X = 0$, dann bedeutet es $\text{E}(X-\text{E}X)^2 = 0$ für $(X-\text{E}X)^2 \geq 0$. Damit folgt nach Satz 4.1.1, 8) $(X-\text{E}X)^2 = 0$ fast sicher $\Longrightarrow X = \text{E}X$ fast sicher.

4. Es gilt

$$\begin{aligned}&\text{Var}\,(X+Y) \\ &= \text{E}(X+Y)^2 - (\text{E}(X+Y))^2 \\ &= \text{E}(X^2+2XY+Y^2) - (\text{E}X+\text{E}Y)^2 \\ &= \text{E}(X^2) + 2\text{E}(XY) + \text{E}Y^2 - (\text{E}X)^2 - 2\text{E}X\cdot\text{E}Y - (\text{E}Y)^2 \\ &= \underbrace{\text{E}(X^2)-(\text{E}X)^2}_{\text{Var}\,X} + \underbrace{\text{E}(Y^2)-(\text{E}Y)^2}_{\text{Var}\,Y} + \underbrace{2(\text{E}(XY)-\text{E}X\cdot\text{E}Y)}_{\text{Cov}\,(X,Y)} \\ &= \text{Var}\,X + \text{Var}\,Y + 2\text{Cov}\,(X,Y)\,.\end{aligned}$$

5. Falls X und Y unabhängig sind, dann gilt nach dem Satz 4.1.1, 7) $\text{E}(XY) = \text{E}X\cdot\text{E}Y$ und somit $\text{Cov}\,(X,Y) = \text{E}(XY) - \text{E}X\cdot\text{E}Y = 0$.

□

Folgerung 4.2.1 1. Es gilt $\text{Var}\,a = 0 \quad \forall a \in \mathbb{R}$.

2. Falls $\text{Var}\,X = 0$, dann ist $X = const$ fast sicher.

3. Für Zufallsvariablen $X_1, \ldots, X_n$ mit $\text{E}X_i^2 < \infty$, $i = 1, \ldots, n$ gilt

$$\text{Var}\,\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}\,X_i + 2\sum_{i<j}\text{Cov}\,(X_i, X_j)\,.$$

4. Falls $X_1, \ldots, X_n$ paarweise unkorreliert sind, dann gilt

$$\text{Var}\,\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}\,X_i\,.$$

Dies gilt insbesondere dann, wenn die Zufallsvariablen $X_1, \ldots, X_n$ paarweise unabhängig sind.

Beispiel 4.2.1

1. *Poisson–Verteilung:*
   Sei $X \sim Poisson(\lambda)$, $\lambda > 0$. Zeigen wir, dass Var $X = \lambda$.
   Es ist uns bereits bekannt, dass $\mathrm{E}X = \lambda$. Somit gilt

$$\begin{aligned}
\text{Var } X &= \mathrm{E}(X^2) - \lambda^2 = \sum_{k=0}^{\infty} k^2 \underbrace{e^{-\lambda}\frac{\lambda^k}{k!}}_{=P(X=k)} - \lambda^2 \\
&= e^{-\lambda}\sum_{k=1}^{\infty}(k(k-1)+k)\frac{\lambda^k}{k!} - \lambda^2 \\
&= e^{-\lambda}\sum_{k=1}^{\infty} k(k-1)\frac{\lambda^k}{k!} + \underbrace{\sum_{k=1}^{\infty} k e^{-\lambda}\frac{\lambda^k}{k!}}_{=\mathrm{E}X=\lambda} - \lambda^2 \\
&= e^{-\lambda}\sum_{k=2}^{\infty}\frac{\lambda^{k-2}\cdot\lambda^2}{(k-2)!} + \lambda - \lambda^2 \\
&\underbrace{=}_{m=k-2} \lambda^2 \underbrace{\sum_{m=0}^{\infty} e^{-\lambda}\cdot\frac{\lambda^m}{m!}}_{=1} + \lambda - \lambda^2 = \lambda\,.
\end{aligned}$$

2. *Normalverteilung:*
   Sei $X \sim N(\mu, \sigma^2)$. Zeigen wir, dass Var $X = \sigma^2$. Wie wir wissen, gilt $\mathrm{E}X = \mu$ und somit

$$\begin{aligned}
\text{Var } X &= \mathrm{E}(X-\mu)^2 = \int_{\mathbb{R}} (x-\mu)^2 \underbrace{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}_{=f_X(x)}\, dx \\
&\underbrace{=}_{y=\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}}\sigma^2 \int_{\mathbb{R}} y^2 e^{-\frac{y^2}{2}}\, dy \\
&= \frac{1}{\sqrt{2\pi}}\sigma^2 \int_{-\infty}^{\infty} y e^{-\frac{y^2}{2}}\, d\left(\frac{y^2}{2}\right) = \frac{-1}{\sqrt{2\pi}}\sigma^2 \int_{-\infty}^{\infty} y\, d\left(e^{-\frac{y^2}{2}}\right) \\
&= \frac{1}{\sqrt{2\pi}}\sigma^2 \left(-y e^{-\frac{y^2}{2}}\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}}\, dy \right) \\
&= \frac{1}{\sqrt{2\pi}}\sigma^2 \left(-0 + \sqrt{2\pi}\right) = \sigma^2\,.
\end{aligned}$$

4.3 Kovarianz und Korrelationskoeffizient

Definition 4.3.1 Seien X und Y zwei Zufallsvariablen mit $0 < \text{Var } X$, $\text{Var } Y < \infty$. Die Größe

$$\varrho(X,Y) = \frac{\text{Cov }(X,Y)}{\sqrt{\text{Var } X}\sqrt{\text{Var } Y}}$$

heißt *Korrelationskoeffizient* von X und Y.
$\varrho(X,Y)$ ist ein Maß für die lineare Abhängigkeit der Zufallsvariablen X und Y.

Satz 4.3.1 *(Eigenschaften des Korrelationskoeffizienten):*
Seien X und Y zwei Zufallsvariablen mit $0 < \text{Var } X$, $\text{Var } Y < \infty$. Dann gilt

1. $|\varrho(X,Y)| \le 1$.
2. $\varrho(X,Y) = 0$ genau dann, wenn X und Y unkorreliert sind. Eine hinreichende Bedingung dafür ist die Unabhängigkeit von X und Y.
3. $|\varrho(X,Y)| = 1$ genau dann, wenn X und Y fast sicher linear abhängig sind, d.h., $\exists a \neq 0,\ b \in \mathbb{R}:\ P(Y = aX + b) = 1$.

Beweis Setzen wir

$$\bar{X} = \frac{X - \text{E}X}{\sqrt{\text{Var } X}}, \quad \bar{Y} = \frac{Y - \text{E}Y}{\sqrt{\text{Var } Y}}.$$

Diese Konstruktion führt zu den sogenannten *standardisierten Zufallsvariablen* $\bar{X}$ und $\bar{Y}$, für die

$$\begin{aligned} &\text{E}\bar{X} = 0, \quad \text{Var } \bar{X} = 1, \quad \text{Cov }(\bar{X},\bar{Y}) = \text{E}(\bar{X}\bar{Y}) = \varrho(X,Y) \\ &\text{E}\bar{Y} = 0, \quad \text{Var } \bar{Y} = 1. \end{aligned}$$

1. Es gilt

$$\begin{aligned} 0 \ &\le \ \text{Var }(\bar{X} \pm \bar{Y}) = \text{E}(\bar{X} \pm \bar{Y})^2 = \underbrace{\text{E}(\bar{X})^2}_{\text{Var } \bar{X}=1} \pm 2\text{E}(\bar{X}\cdot\bar{Y}) + \underbrace{\text{E}(\bar{Y})^2}_{\text{Var } \bar{Y}=1} \\ &= \ 2 \pm 2\varrho(X,Y) \Longrightarrow 1 \pm \varrho(X,Y) \ge 0 \Longrightarrow |\varrho(X,Y)| \le 1. \end{aligned}$$

2. Folgt aus der Definition 4.3.1 und dem Satz 4.2.1, 5).

3. "$\Leftarrow$" Falls $Y = aX + b$ fast sicher, $a \neq 0$, $b \in \mathbb{R}$, dann gilt Folgendes: Bezeichnen wir $\text{E}X = \mu$ und $\text{Var } X = \sigma^2$. Dann ist $\text{E}Y = a\mu + b$, $\text{Var } Y = a^2 \cdot \sigma^2$ und somit

$$\begin{aligned} \varrho(X,Y) = \text{E}(\bar{X}\bar{Y}) &= \text{E}\left(\frac{X-\mu}{\sigma} \cdot \frac{aX + b - a\mu - b}{|a| \cdot \sigma}\right) \\ &= \text{E}\Bigg(\underbrace{\left(\frac{X-\mu}{\sigma}\right)^2}_{\bar{X}^2} \cdot \text{sgn } a\Bigg) = \text{sgn } a \cdot \underbrace{\text{E}(\bar{X}^2)}_{\text{Var } \bar{X}=1} = \text{sgn } a = \pm 1. \end{aligned}$$

"⇒" Sei $|\varrho(X,Y)| = 1$. O.B.d.A. betrachten wir den Fall $\varrho(X,Y) = 1$. Aus 1) gilt Var $(\bar{X} - \bar{Y}) = 2 - 2\varrho(X,Y) = 0 \Longrightarrow \bar{X} - \bar{Y} = const$ fast sicher aus dem Satz 4.2.1, 3). Somit sind X und Y linear abhängig.
Für den Fall $\varrho(X,Y) = -1$ betrachten wir analog

$$\text{Var } (\bar{X} + \bar{Y}) = 2 + 2\varrho(X,Y) = 0\,.$$

□

4.4 Höhere und gemischte Momente

Außer des Erwartungswertes, der Varianz und der Kovarianz gibt es eine Reihe von weiteren Charakteristiken von Zufallsvariablen, die für uns von Interesse sind.

Definition 4.4.1 Seien $X, X_1, \ldots, X_n$ Zufallsvariablen auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, P)$.

1. Der Ausdruck $\mu_k = \mathrm{E}(X^k)$, $k \in \mathbb{N}$ heißt *k–tes Moment* der Zufallsvariablen X.
2. $\bar{\mu}_k = \mathrm{E}(X - \mathrm{E}X)^k$, $k \in \mathbb{N}$ heißt *k–tes zentriertes Moment* der Zufallsvariablen X.
3. $\mathrm{E}\left(X_1^{k_1} \cdot \ldots \cdot X_n^{k_n}\right)$, $k_1, \ldots, k_n \in \mathbb{N}$ heißt *gemischtes Moment* der Zufallsvariablen $X_1, \ldots, X_n$ der Ordnung $k = k_1 + \ldots + k_n$.
4. $\mathrm{E}\left[(X_1 - \mathrm{E}X_1)^{k_1} \cdot \ldots \cdot (X_n - \mathrm{E}X_n)^{k_n}\right]$ heißt *zentriertes gemischtes Moment* der Zufallsvariablen $X_1, \ldots, X_n$ der Ordnung $k = k_1 + \ldots + k_n$.

Anmerkung:

(a) Die angegebenen Momente müssen nicht unbedingt existieren, beispielsweise existiert $\mathrm{E}X^k$, $k \in \mathbb{N}$ für $X \sim Cauchy(0,1)$ nicht.

(b) Offensichtlich ist Var X das zweite zentrierte Moment von X, genauso wie Cov (X,Y) das zweite zentrierte gemischte Moment von X und Y ist. Dabei haben Momente dritter und vierter Ordnung eine besondere Bedeutung:

Definition 4.4.2 1. Der Quotient

$$\gamma_1 = Sch(X) = \frac{\bar{\mu}_3}{\sqrt{(\bar{\mu}_2)^3}} = \frac{\mathrm{E}(X - \mathrm{E}X)^3}{\sqrt{(\mathrm{Var}\ X)^3}} = \mathrm{E}\left(\bar{X}^3\right)$$

heißt *Schiefe* oder *Symmetriekoeffizient* der Verteilung von X. Falls $\gamma_1 > 0$, dann ist die Verteilung von X rechtsschief bzw. linkssteil (für

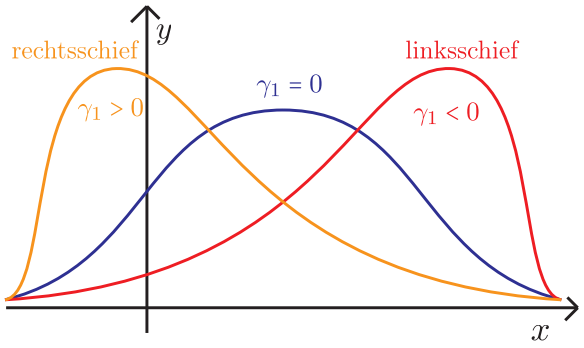


Abbildung 4.1: Veranschaulichung der Schiefe einer Verteilung an Hand der Grafik ihrer Dichte

$\gamma_1 < 0$ linksschief bzw. rechtssteil) vgl. hierzu Abbildung 4.1. Es ist ein Maß für die Symmetrie der Verteilung.

2. Der Ausdruck

$$\gamma_2 = \frac{\bar{\mu}_4}{\bar{\mu}_2^2} - 3 = \frac{\mathrm{E}(X - \mathrm{E}X)^4}{(\mathrm{Var}\ X)^2} - 3 = \mathrm{E}\left(\bar{X}^4\right) - 3$$

heißt *Wölbung (Exzess)* der Verteilung von X. Es ist ein Maß für die "Spitzigkeit" der Verteilung:

$\gamma_2 > 0$ – Verteilung steilgipflig
$\gamma_2 < 0$ – Verteilung flachgipflig, vgl. Abb. 4.2.

Die beiden Kerngrößen messen Abweichungen der Verteilung von X

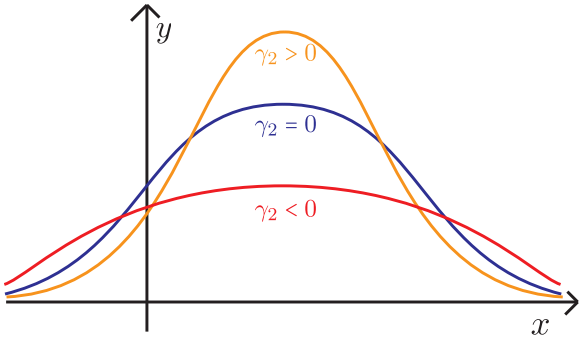


Abbildung 4.2: Veranschaulichung der Wölbung einer Verteilung an Hand der Grafik ihrer Dichte

von einer Gaußschen $N(\mu, \sigma^2)$–Verteilung, für die $\gamma_1 = \gamma_2 = 0$.

Übungsaufgabe 4.4.1 Beweisen Sie, dass $\gamma_1 = \gamma_2 = 0$ für $X \sim N(\mu, \sigma^2)$.

4.5 Entropie

In der Physik wird die *Entropie* als ein logarithmisches Maß für das Chaos bzw. Ordnung, Diversität oder Vielfalt der energetischen Ebenen eines thermodynamischen Systems verstanden. In der Wahrscheinlichkeitstheorie ist es schlicht eine weitere Erwartung–basierte Charakteristik einer Zufallsvariablen X, die die Größe des Wertebereichs C von X bzw. ihre Konzentration innerhalb von C wiedergibt.

Nachdem C. Shannon (1943-1948) den von L. Boltzmann in den 1870er Jahren eingeführten statistischen Begriff der Entropie in der Informationstheorie anwendete, wurde dieser in den Arbeiten [26, 27] von A. Rényi weiter verallgemeinert:

Definition 4.5.1 Sei die Zufallsvariable $X : \Omega \to \mathbb{R}$ diskret verteilt mit Wertebereich C.

1. Die *Shannon*[1]*–Entropie* von X wird eingeführt als
$$H(X) := -\sum_{x \in C} \Pr(X = x) \log \Pr(X = x)\,.$$

2. Die *Rényi*[2]*–Entropie* der Ordnung $\alpha > 0,\ \alpha \neq 1$ von X wird definiert durch
$$H_\alpha(X) := (1-\alpha)^{-1} \log \left(\sum_{x \in C} (\Pr(X = x))^\alpha \right)\,.$$

Definition 4.5.2 Die Zufallsvariable $X : \Omega \to \mathbb{R}$ sei absolut stetig verteilt mit Dichte f_X.

1. Die *differentielle Entropie* von X nennt man die Größe
$$h(X) := -\int_{\mathbb{R}} f_X(y) \log f_X(y)\, dy.$$

2. Die *differentielle Rényi–Entropie* der Ordnung $\alpha > 0,\ \alpha \neq 1$ von X ist entsprechend gleich
$$h_\alpha(X) := (1-\alpha)^{-1} \log \int_{\mathbb{R}} f_X^\alpha(y)\, dy\,.$$

Auf Grund des folgenden Grenzwertverhaltens für $\alpha \to 1$ werden wir die Bezeichnungen $H_1 = H$, $h_1 = h$ verwenden:

[1] Benannt nach dem amerikanischen Mathematiker Claude E. Shannon (1916-2001), dem Vater der stochastischen Informationstheorie.

[2] Benannt nach ihrem Urheber ungarischen Mathematiker Alfréd Rényi (1921-1970).

Lemma 4.5.1 (Eigenschaften der Entropie) Sei X eine Zufallsvariable mit einer diskreten bzw. absolut stetigen Verteilung. Für beliebiges $\alpha > 0$ gelten folgende Eigenschaften:

1. Asymptotik in α: $H(X) = \lim_{\alpha\to 1} H_\alpha(X)$ bzw. $h(X) = \lim_{\alpha\to 1} h_\alpha(X)$.
2. Verschiebungsinvarianz: $H_\alpha(X+b) = H_\alpha(X)$ bzw. $h_\alpha(X+b) = h_\alpha(X)$ für beliebiges $b \in \mathbb{R}$.
3. Skalierung: $H_\alpha(aX) = H_\alpha(X)$, $h_\alpha(aX) = h_\alpha(X) + \log|a|$ für beliebiges $a \neq 0$.

Beweis

1. Beide Aussagen folgen unter Verwendung von der L'Hospital-Regel, indem man den Zähler in H_α, h_α und den Nenner $1-\alpha$ bzgl. α differenziert und dann $\alpha \to 1$ streben lässt.
2. Für diskrete Zufallsvariablen X bewirkt die Addition einer Konstanten b lediglich die Verschiebung ihres Wertebereichs $(C+b)$, was durch die Substitution $x \mapsto x - b$ in der Definition 4.5.1 behoben wird. Für absolut stetig verteilte Zufallsvariablen X folgt die Aussage direkt aus Satz 3.6.2.
3. Beweisen wir die Aussagen für $\alpha \neq 1$. Der Fall $\alpha = 1$ folgt daraus mit Hilfe des Grenzwertübergangs für $\alpha \to 1$, siehe Punkt 1. Die Behauptung für H_α wird analog wie im Punkt 2 bewiesen mit Substitution $x \mapsto x/a$ in der Summe. Nach dem Satz 3.6.2 gilt für h_α, dass

$$\begin{aligned} h_\alpha(aX) &= (1-\alpha)^{-1} \log \int_{\mathbb{R}} |a|^{-\alpha} f_X^\alpha(y/a)\, dy \\ &= (1-\alpha)^{-1} \log |a|^{1-\alpha} + (1-\alpha)^{-1} \log \int_{\mathbb{R}} f_X^\alpha(y/a)\, d(y/a) \\ &= \log|a| + h_\alpha(X). \end{aligned}$$

□

Bemerkung 4.5.1

1. Es gilt

$$H(X) = -\mathbb{E}\log p_X(X), \quad H_\alpha(X) = (1-\alpha)^{-1}\log \mathbb{E} p_X^{\alpha-1}(X),$$

$$h(X) = -\mathbb{E}\log f_X(X), \quad h_\alpha(X) = (1-\alpha)^{-1}\log \mathbb{E} f_X^{\alpha-1}(X),$$

wobei p_X bzw. f_X die (Zähl)Dichte der diskret bzw. absolut stetig verteilten Zufallsvariablen X ist. Dei erste dieser Gleichungen wird in der Informationstheorie als die erwartete Information interpretiert, gewonnen aus der Beobachtung von X.

2. Die Entropien H_α bzw. h_α, $\alpha > 0$, können auf ähnlichem Wege auch für Zufallsvektoren (H_α sogar für beliebige Zufallselemente) eingeführt werden, da ihre Definition lediglich von den (Zähl)Dichten Gebrauch macht.

3. Sei $\alpha > 0$ beliebig. Obwohl $H_\alpha(X) \geq 0$ für alle diskret verteilten Zufallsvariablen X, kann $h_\alpha(X)$ auch negative Werte annehmen, vgl. Beispiel 4.5.1.

4. In der Informationstheorie ist es üblich, den natürlichen Logarithmus (zur Basis e) in den Definitionen 4.5.1, 4.5.2 durch den $\log_2$ zu ersetzen, was mit der binären Kodierung bei der Informationsübertragung zusammenhängt.

Beispiel 4.5.1 Die Entropie H_α bzw. h_α von zwei wichtigen Verteilungen sei hier gegeben, wobei $\alpha > 0$ beliebig ist:

1. Gleichverteilung: Für $X \sim \mathcal{U}\{x_1, \ldots, x_n\}$, $n \in \mathbb{N}$, und $Y \sim \mathcal{U}[0, M]$, $M > 0$, gilt $H_\alpha(X) = \log n$, $h_\alpha(Y) = \log M$. Dabei ist die Entropie gleich Null für eine konstante X und tendiert gegen $+\infty$ mit unbegrenzt wachsender Anzahl n der Zustände des Systems. Für $M \to +0$ gilt dagegen $h_\alpha(Y) \to -\infty$.

2. Normalverteilung: Für $X \sim N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma > 0$, gilt

$$h_\alpha(X) = \log \sigma + \frac{1}{2} \log 2\pi + \begin{cases} \frac{1}{2}, & \alpha = 1\,, \\ \frac{\log \alpha}{2(\alpha-1)}, & \alpha > 0, \alpha \neq 1\,. \end{cases}$$

   Wir schlussfolgern $\lim_{\sigma \to +0} h_\alpha(X) = -\infty$, $\lim_{\sigma \to +\infty} h_\alpha(X) = +\infty$.

Übungsaufgabe 4.5.1 Zeigen Sie, dass die Entropie einer Zufallsvariablen $X \sim Exp(\lambda)$, $\lambda > 0$, durch $h_\alpha(X) = \frac{\lambda^{\alpha-1}}{(1-\alpha)\alpha}$, $\alpha > 0$, $\alpha \neq 1$, sowie $h(X) = 1 - \log \lambda$ gegeben ist.

Die statistische Schätzung der Entropie wird in den Arbeiten [15, 4, 33, 6, 22, 19, 18, 25] ausführlich behandelt.

4.6 Ungleichungen

Satz 4.6.1 *(Ungleichung von Markow):*
Sei X eine Zufallsvariable mit $\mathrm{E}|X|^r < \infty$ für ein $r \geq 1$. Dann gilt

$$P(|X| \geq \varepsilon) \leq \frac{\mathrm{E}|X|^r}{\varepsilon^r} \quad \forall \varepsilon > 0\,.$$

Beweis Es gilt

$$\mathrm{E}|X|^r = \underbrace{\mathrm{E}\left(|X|^r \cdot I(|X| \leq \varepsilon)\right)}_{\geq 0} + \mathrm{E}\left(|X|^r \cdot I(|X| > \varepsilon)\right)$$

$$\geq \quad \mathrm{E}\left(\varepsilon^r \cdot I(|X| > \varepsilon)\right) = \varepsilon^r \cdot P(|X| > \varepsilon)\,,$$

daher $P(|X| > \varepsilon) \leq \frac{\mathrm{E}|X|^r}{\varepsilon^r}$. □

Folgerung 4.6.1 *(Ungleichung von Tschebyschew).*

1. Sei X eine Zufallsvariable mit $\mathrm{E}X^2 < \infty$ und $\varepsilon > 0$. Dann gilt
$$P(|X - \mathrm{E}X| \geq \varepsilon) \leq \frac{\mathrm{Var}\ X}{\varepsilon^2}\,.$$
2. Sei $\mathrm{E}e^{sX} < \infty$ für ein $s > 0$. Dann gilt
$$P(X \geq \varepsilon) \leq \frac{\mathrm{E}e^{\lambda X}}{e^{\lambda\varepsilon}}\,, \forall\varepsilon > 0 \quad \forall 0 \leq \lambda \leq s\,.$$

Beweis Benutze die Markow–Ungleichung für die Zufallsvariable

1. $Y = X - \mathrm{E}X$ und $r = 2$ und
2. $Y = e^{\lambda X} \geq 0$, $\varepsilon = e^{\lambda\varepsilon_0}$, $r = 1$.

□

Beispiel 4.6.1 Der Durchmesser der Mondscheibe wird aus den Bildern der Astrophotographie folgendermaßen bestimmt: bei jeder Aufnahme der Mondscheibe wird ihr Durchmesser X_i gemessen. Nach n Messungen wird der Durchmesser als $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ aus allen Beobachtungen geschätzt. Sei $\mu = \mathrm{E}X_i$ der wahre (unbekannte) Wert des Monddurchmessers. Wie viele Beobachtungen n müssen durchgeführt werden, damit die Schätzung $\bar{X}_n$ weniger als um $0,1$ vom Wert μ mit Wahrscheinlichkeit von mindestens $0,99$ abweicht? Mit anderen Worten, finde n: $P(|\bar{X}_n - \mu| \leq 0,1) \geq 0,99$. Diese Bedingung ist äquivalent zu $P(|\bar{X}_n - \mu| > 0,1) \leq 1 - 0,99 = 0,01$. Sei $\mathrm{Var}\ X_i = \sigma^2 > 0$. Dann gilt

$$\mathrm{Var}\ \bar{X}_n = \frac{1}{n^2} \cdot \sum_{i=1}^n \mathrm{Var}\ X_i = \frac{1}{n^2} \cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}\,,$$

wobei vorausgesetzt wird, dass alle Messungen X_i unabhängig voneinander durchgeführt werden. Somit gilt nach der Ungleichung von Tschebyschew

$$P(|\bar{X}_n - \mu| > 0,1) \leq \frac{\mathrm{Var}\ \bar{X}_n}{0,1^2} = \frac{\sigma^2}{n \cdot 0,01}\,,$$

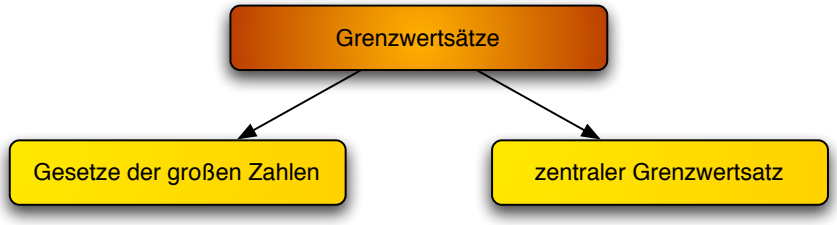
woraus folgt, dass

$$n \geq \frac{\sigma^2}{(0,01)^2} = 10^4 \cdot \sigma^2\,.$$

Für $\sigma = 1$ braucht man z.B. mindestens 10000 Messungen! Diese Zahl zeigt, wie ungenau die Schranke in der Ungleichung von Tschebyschew ist. Eine viel genauere Antwort ($n \geq 670$) kann man mit Hilfe des zentralen Grenzwertsatzes bekommen. Dies wird allerdings erst im Kapitel 5 behandelt.

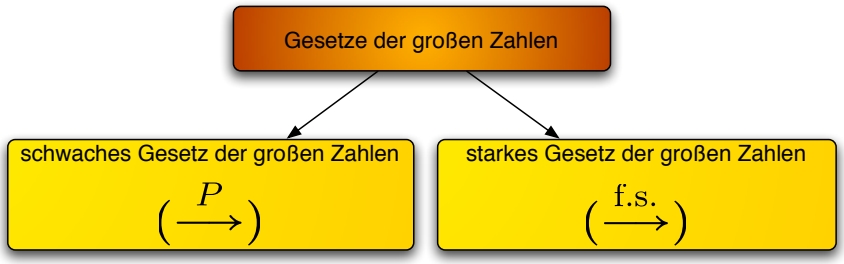
Kapitel 5

Grenzwertsätze



In diesem Kapitel betrachten wir Aussagen der Wahrscheinlichkeitsrechnung, die Näherungsformeln von großer anwendungsbezogener Bedeutung liefern. Dies wird an mehreren Beispielen erläutert.

5.1 Gesetze der großen Zahlen



Ein typisches Gesetz der großen Zahlen besitzt die Form

$$\frac{1}{n}\sum_{i=1}^{n} X_i \underset{n\to\infty}{\longrightarrow} \mathrm{E}X_0\,, \tag{5.1}$$

wobei $\{X_n\}_{n\in\mathbb{N}}$ unabhängige identisch verteilte Zufallsvariablen mit $X_n \stackrel{d}{=} X_0$, $\mathrm{E}|X_0| < \infty$ sind.
Die Konvergenz in (5.1) wird entweder in Wahrscheinlichkeit oder fast sicher verstanden.

Definition 5.1.1 Sei $\{X_n\}_{n\in\mathbb{N}}$ eine Folge von Zufallsvariablen definiert auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, P)$. Sei X eine weitere Zufallsvariable auf $(\Omega, \mathcal{F}, P)$. Man sagt, die Folge $\{X_n\}$ konvergiert gegen X für $n \to \infty$

1. *fast sicher oder mit Wahrscheinlichkeit* 1 ($X_n \xrightarrow[n\to\infty]{\text{f.s.}} X$), falls
$$P(\{\omega \in \Omega : X_n(\omega) \underset{n\to\infty}{\longrightarrow} X(\omega)\}) = 1\,.$$

2. *in Wahrscheinlichkeit oder stochastisch* ($X_n \xrightarrow[n\to\infty]{P} X$), falls
$$\forall \varepsilon > 0 \quad P(|X_n - X| > \varepsilon) \underset{n\to\infty}{\longrightarrow} 0\,.$$

Wenn $\xrightarrow{P}$ gemeint ist, spricht man von dem *schwachen Gesetz der großen Zahlen*. Falls $\xrightarrow{\text{f.s.}}$ gemeint ist, heißt die Aussage (5.1) *starkes Gesetz der großen Zahlen*.

Im Folgenden verwenden wir die Bezeichnungen $S_n = \sum_{i=1}^n X_i$, $\overline{X}_n = \frac{S_n}{n}$ für all $n \in \mathbb{N}$, für eine Folge $\{X_n\}_{n\in\mathbb{N}}$ von Zufallsvariablen auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, P)$.

5.1.1 Schwaches und Starkes Gesetz der großen Zahlen

Satz 5.1.1 *(Markow)*
Sei $\{X_n\}_{n\in\mathbb{N}}$ eine Folge von Zufallsvariablen auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, P)$ mit $\mathrm{E}X_i^2 < \infty\, \forall i$. Falls
$$\mathrm{Var}\ \overline{X}_n \underset{n\to\infty}{\longrightarrow} 0\,, \tag{5.2}$$
dann gilt
$$\overline{X}_n - \frac{1}{n}\sum_{i=1}^n \mathrm{E}X_i \xrightarrow[n\to\infty]{P} 0\,.$$

Folgerung 5.1.1 Seien die Zufallsvariablen $\{X_n\}_{n\in\mathbb{N}}$ im Satz 5.1.1 unabhängig. Dann gilt Folgendes:

1. Die Bedingung $\mathrm{Var}\ \overline{X}_n \underset{n\to\infty}{\longrightarrow} 0$ bekommt die Form
$$\frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}\ X_i \underset{n\to\infty}{\longrightarrow} 0\,.$$

2. Falls Var $X_n \leq c = const \quad \forall n \in \mathbb{N}$, dann gilt die Bedingung (5.2) und somit die Aussage des Satzes 5.1.1 (Satz von Tschebyschew).

3. Insbesondere ist die Bedingung Var $X_n \leq c = const \quad \forall n \in \mathbb{N}$ erfüllt, falls $\{X_n\}_{n\in\mathbb{N}}$ unabhängige identisch verteilte Zufallsvariablen mit

$$\mathrm{E}X_n = \mu, \text{ Var } X_n = \sigma^2 < \infty$$

sind. Dann nimmt das schwache Gesetz der großen Zahlen die klassische Form

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{P} \mu$$

an.

Die Existenz der zweiten Momente ist für das schwache Gesetz der großen Zahlen nicht entscheidend. So kann man mit Hilfe der charakteristischen Funktionen folgenden Satz beweisen:

Satz 5.1.2 *(Schwaches Gesetz der großen Zahlen von Kchintschin)*
Sei $\{X_n\}_{n\in\mathbb{N}}$ eine Folge von stochastisch unabhängigen, integrierbaren Zufallsvariablen, $n \in \mathbb{N}$, mit demselben Erwartungswert $\mathrm{E}X_n = \mu < \infty$. Dann gilt

$$\overline{X}_n \xrightarrow[n\to\infty]{P} \mu\,.$$

Satz 5.1.3 *(Starkes Gesetz der großen Zahlen von Kolmogorow)*
Seien $\{X_n\}_{n\in\mathbb{N}}$ unabhängige identisch verteilte Zufallsvariablen. Es gilt, dass $\overline{X}_n \xrightarrow[n\to\infty]{\text{f.s.}} \mu$ genau dann, wenn $\exists\, \mathrm{E}X_n = \mu < \infty$.

5.1.2 Anwendung der Gesetze der großen Zahlen

1. *Monte–Carlo–Methoden zur numerischen Integration*
Sei $g : [0,1]^d \longrightarrow \mathbb{R}$ eine beliebige stetige Funktion. Wie kann man mit Hilfe der Gesetze der großen Zahlen

$$\int_{[0,1]^d} g(x)\,dx = \int_0^1 \ldots \int_0^1 g(x_1,\ldots,x_d)\,dx_1 \ldots\, dx_d$$

numerisch berechnen?

Der Algorithmus ist wie folgt:

- Generiere eine Folge von Realisierungen von unabhängigen identisch verteilten Zufallsvariablen

$$X_1,\ldots,X_n \text{ mit } X_i \sim [0,1]^d,\ i = 1,\ldots,n\,.$$

- Setze

$$\int_{[0,1]^d} g(x)\,dx \approx \frac{1}{n}\sum_{i=1}^{n} g(X_i) \tag{5.3}$$

für große n. Dieser Vorgang ist berechtigt, denn nach dem Satz 5.1.3 gilt

$$\frac{1}{n}\sum_{n=1}^{n} g(X_i) \xrightarrow[n\to\infty]{\text{f.s.}} \mathrm{E}g(X_1) = \int_{[0,1]^d} g(x)\,dx$$

und somit gilt (5.3) für ausreichend große n.

Bemerkung:
Dieselbe Methode kann durch geeignete Transformation vom Integrationsgebiet $G \subset \mathbb{R}^d$ und andere Wahl von Zufallsvariablen X_i auf die Berechnung von $\int_G g(x)\,dx$ erweitert werden, G kompakte Teilmenge von $\mathbb{R}^d$. So genügt es nur $X_i \sim U(G)$, $i = 1, \ldots, n$ zu betrachten.

2. *Numerische Berechnung der Zahl π:*
Wie kann π mit Hilfe eines Rechners beliebig genau berechnet werden? Dazu wird das starke Gesetz der großen Zahlen wie folgt verwendet:

- Generiere Realisierungen von unabhängig und identisch verteilten Zufallvektoren $X_1, \ldots, X_n \in \mathbb{R}^2$ mit $X_i \sim U[-1,1]^2$, $i = 1, \ldots, n$.
- Es gilt

$$\pi \approx \frac{4}{n}\sum_{i=1}^{n} I(|X_i| \le 1) \tag{5.4}$$

für große n.

In der Tat, nach dem Satz 5.1.3 gilt

$$\frac{1}{n}\sum_{i=1}^{n} I(|X_i| \le 1) \xrightarrow[n\to\infty]{\text{f.s.}} \mathrm{E}I(|X_1| \le 1) = P(|X_1| \le 1) = \frac{|B_1(0)|}{|[-1,1]^2|} = \frac{\pi}{2^2} = \frac{\pi}{4}.$$

Somit ist die Verwendung der Berechnungsformel (5.4) berechtigt für große n.

5.2 Zentraler Grenzwertsatz

Für die Gesetze der großen Zahlen wurde die Normierung $\frac{1}{n}$ der Summe $S_n = \sum_{i=1}^n X_i$ gewählt, um $\frac{S_n}{n} \underset{n\to\infty}{\longrightarrow} \mathrm{E}X_1$ zu beweisen. Falls jedoch eine andere Normierung gewählt wird, so sind andere Grenzwertaussagen möglich. Im Fall der Normierung $\frac{1}{\sqrt{n}}$ spricht man von zentralen Grenzwertsätzen: unter gewissen Voraussetzungen gilt also

$$\frac{S_n - n\mathrm{E}X_1}{\sqrt{n\mathrm{Var}\ X_1}} \underset{n\to\infty}{\overset{d}{\longrightarrow}} Y \sim N(0,1)\,.$$

5.2.1 Klassischer zentraler Grenzwertsatz

Satz 5.2.1 Sei $\{X_n\}_{n\in\mathbb{N}}$ eine Folge von unabhängigen identisch verteilten Zufallsvariablen auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, P)$ mit $\mathrm{E}X_i = \mu$, $\mathrm{Var}\ X_i = \sigma^2$, wobei $0 < \sigma^2 < \infty$. Dann gilt

$$P\left(\frac{S_n - n\mu}{\sqrt{n}\sigma} \le x\right) \underset{n\to\infty}{\longrightarrow} \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-y^2}{2}}\, dy \quad \forall x \in \mathbb{R}\,,$$

wobei $\Phi(x)$ die Verteilungsfunktion einer $N(0,1)$–verteilten Zufallsvariablen ist.

Folgerung 5.2.1 Unter den Voraussetzungen des Satzes 5.2.1 gilt zusätzlich

1. $$P\left(\frac{S_n - n\mu}{\sqrt{n}\cdot\sigma} < x\right) \underset{n\to\infty}{\longrightarrow} \Phi(x) \quad \forall x \in \mathbb{R}\,,$$

2. $$P\left(a \le \frac{S_n - n\mu}{\sqrt{n}\cdot\sigma} \le b\right) \underset{n\to\infty}{\longrightarrow} \Phi(b) - \Phi(a) \quad \forall a, b \in \mathbb{R},\ a \le b\,.$$

Beispiel 5.2.1 1. *Satz von de Moivre–Laplace*
Falls $X_n \sim Bernoulli(p)$, $n \in \mathbb{N}$ unabhängig sind und $p \in (0,1)$, dann genügt die Folge $\{X_n\}_{n\in\mathbb{N}}$ mit $\mathrm{E}X_n = p$, $\mathrm{Var}\ X_n = p(1-p)$ den Voraussetzungen des Satzes 5.2.1. Das Ergebnis

$$\frac{S_n - np}{\sqrt{np(1-p)}} \underset{n\to\infty}{\overset{d}{\longrightarrow}} Y \sim N(0,1)$$

wurde mit einfachen Mitteln als erster zentraler Grenzwertsatz von Abraham de Moivre (1667-1754) bewiesen und trägt daher seinen Namen. Es kann folgendermaßen interpretiert werden:

Falls die Anzahl n der Experimente groß wird, so wird die Binomialverteilung von $S_n \sim Bin(n,p)$, das die Anzahl der Erfolge in n Experimenten darstellt, approximiert durch

$$P(a \leq S_n \leq b) = P\left(\frac{a-np}{\sqrt{np(1-p)}} \leq \frac{S_n - np}{\sqrt{p(1-p)}} \leq \frac{b-np}{\sqrt{np(1-p)}}\right)$$
$$\approx \Phi\left(\frac{b-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a-np}{\sqrt{np(1-p)}}\right),$$

$\forall\, a, b \in \mathbb{R}$, $a \leq b$, wobei $p \in (0,1)$ als die Erfolgswahrscheinlichkeit in einem Experiment interpretiert wird. So kann z.B. S_n als die Anzahl von Kopf in n Würfen einer fairen Münze ($p = \frac{1}{2}$) betrachtet werden. Hier gilt also

$$P(a \leq S_n \leq b) \underset{n\text{–groß}}{\approx} \Phi\left(\frac{2b-n}{\sqrt{n}}\right) - \Phi\left(\frac{2a-n}{\sqrt{n}}\right), \quad a < b\,.$$

2. Berechnen wir die Anzahl der notwendigen Messungen des Monddurchmessers im Beispiel 4.6.1 mit Hilfe des zentralen Grenzwertsatzes:
Im Allgemeinen gilt es ein $n \in \mathbb{N}$ zu finden, so dass

$$P\left(\left|\overline{X}_n - \mu\right| \leq \varepsilon\right) > 1 - \delta.$$

Für den Fall von Beispiel 4.6.1 folgt

$$P(|\overline{X}_n - \mu| \leq 0,1) > 0,99\,,$$

oder äquivalent dazu

$$P(|\overline{X}_n - \mu| > 0,1) \leq 0,01\,,$$

wobei $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ ist. Es gilt

$$P(|\overline{X}_n - \mu| \leq 0,1) = P\left(-0,1 \leq \frac{S_n - n\mu}{n} \leq 0,1\right)$$
$$= P\left(-0,1\frac{\sqrt{n}}{\sigma} \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq 0,1\frac{\sqrt{n}}{\sigma}\right)$$
$$\underset{n \text{ groß}}{\approx} \Phi\left(\frac{0,1\sqrt{n}}{\sigma}\right) - \Phi\left(-\frac{0,1\sqrt{n}}{\sigma}\right)$$
$$= 2\Phi\left(\frac{0,1\sqrt{n}}{\sigma}\right) - 1 = 2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) - 1 \overset{!}{>} 1 - \delta$$

weil $N(0,1)$ eine symmetrische Verteilung ist und somit

$$\Phi(x) = 1 - \Phi(-x)\,\forall x \in \mathbb{R}$$

gilt.

Es muss also $2\Phi\left(\frac{0,1\sqrt{n}}{\sigma}\right) - 1 > 0,99$ erfüllt sein. Dies ist äquivalent zu

$$\Phi\left(\frac{0,1\sqrt{n}}{\sigma}\right) > \frac{1,99}{2} = 0,995 \quad \Longleftrightarrow \quad \frac{0,1\sqrt{n}}{\sigma} > \Phi^{-1}(0,995)$$

oder

$$\begin{aligned} n &> \frac{\sigma^2}{(0,1)^2}\left(\Phi^{-1}(0,995)^2\right) = \frac{\sigma^2\left(\Phi^{-1}(0,995)\right)^2}{0,01} \\ &= \frac{\sigma^2(2,58)^2}{0,01} = \sigma^2 \cdot 665,64\,. \end{aligned}$$

Für $\sigma^2 = 1$ ergibt sich die Antwort

$$\boxed{n \geq 666}$$

5.2.2 Konvergenzgeschwindigkeit im zentralen Grenzwertsatz

In diesem Abschnitt möchten wir die *Schnelligkeit der Konvergenz im zentralen Grenzwertsatz* untersuchen. Damit aber diese Fragestellung überhaupt sinnvoll erscheint, muss die Konvergenz im zentralen Grenzwertsatz gleichmäßig sein:

$$\sup_{x\in\mathbb{R}}\left|P\left(\frac{\sum_{k=1}^n X_k - n\mu}{\sqrt{n}\sigma} \leq x\right) - \Phi(x)\right| \underset{n\to\infty}{\longrightarrow} 0\,.$$

Das ist tatsächlich der Fall, wie aus der Stetigkeit von $\Phi(x)$ und dem folgenden Lemma 5.2.1 hervorgeht.

Lemma 5.2.1 Seien $\{F_n\}_{n=1}^\infty$, F Verteilungsfunktionen, sodass $F(x)$ stetig ist $\forall x \in \mathbb{R}$ und $F_n(x) \underset{n\to\infty}{\longrightarrow} F(x)$ $\forall x \in \mathbb{R}$. Dann ist die Konvergenz von F_n zu F *gleichmäßig*:

$$\sup_{x\in\mathbb{R}}|F_n(x) - F(x)| \underset{n\to\infty}{\longrightarrow} 0\,.$$

Satz 5.2.2 *(Berry–Esséen)*
Sei $\{X_n\}_{n\in\mathbb{N}}$ eine Folge von unabhängigen identisch verteilten Zufallsvariablen mit $\mathrm{E}X_n = \mu$, $\mathrm{Var}\ X_n = \sigma^2 > 0$, $\mathrm{E}|X_n|^3 < \infty$. Sei

$$F_n(x) = P\left(\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} \leq x\right), \quad x \in \mathbb{R},\, n \in \mathbb{N}\,.$$

Dann gilt

$$\sup_{x\in\mathbb{R}}|F_n(x) - \Phi(x)| \leq \frac{c \cdot \mathrm{E}|X_1 - \mu|^3}{\sigma^3\sqrt{n}}\,,$$

wobei c eine Konstante ist, $\frac{1}{\sqrt{2\pi}} \leq c < 0,4785$, $\frac{1}{\sqrt{2\pi}} \approx 0,39894$.

5.2.3 Grenzwertsatz von Lindeberg

Im klassischen zentralen Grenzwertsatz wurden Folgen von unabhängigen und identisch verteilten Zufallsvariablen betrachtet. In diesem Abschnitt lassen wir die Voraussetzung der identischen Verteiltheit fallen und formulieren einen allgemeineren Grenzwertsatz in der Form von Lindeberg.

Satz 5.2.3 *(Lindeberg)*
Sei $\{X_n\}_{n\in\mathbb{N}}$ eine Folge von unabhängigen Zufallsvariablen mit $\mathrm{E}X_n = \mu_n$, $0 < \sigma_n^2 = \mathrm{Var}\ X_n < \infty\, \forall n$. Sei $S_n = \sum_{i=1}^n X_i$, $D_n^2 = \sum_{i=1}^n \sigma_i^2$. Falls

$$\frac{1}{D_n^2} \sum_{k=1}^{n} \mathrm{E}\left((X_k - \mu_k)^2 \cdot I(|X_k - \mu_k| > \varepsilon D_n)\right) \underset{n\to\infty}{\longrightarrow} 0\,,$$

dann gilt

$$\frac{S_n - \sum_{i=1}^n \mu_i}{D_n} \underset{n\to\infty}{\xrightarrow{d}} Y \sim N(0,1)\,.$$

Folgerung 5.2.2 Sei $\{X_n\}_{n\in\mathbb{N}}$ eine Folge von unabhängigen Zufallsvariablen mit $|X_n| \le c < \infty\, \forall n \in \mathbb{N}$ und $D_n \underset{n\to\infty}{\longrightarrow} \infty$. Dann gilt der zentrale Grenzwertsatz in der Form des Satzes 5.2.3.

Beweis Wir müssen die Gültigkeit der Lindeberg–Bedingung prüfen: aus der Ungleichung von Tschebyschew folgt

$$\begin{aligned}\mathrm{E}\left((X_k - \mu_k)^2 \cdot I(|X_k - \mu_k| > \varepsilon D_n)\right) &\underset{|X_k|\le c, |\mathrm{E}X_k| \le c}{\le} (2c)^2 P(|X_k - \mu_k| > \varepsilon D_n) \\ &\le \quad 4c^2 \frac{\sigma_k^2}{\varepsilon^2 D_n^2} \underset{n\to\infty}{\longrightarrow} 0\,,\end{aligned}$$

für $1 \le k \le n$ und somit

$$\frac{1}{D_n^2} \sum_{k=1}^{n} \mathrm{E}\left((X_k - \mu_k)^2 \cdot I(|X_k - \mu_k| > \varepsilon D_n)\right) \quad \le \quad 4c^2 \frac{\overbrace{\sum_{k=1}^{n} \sigma_k^2}^{=D_n^2}}{\varepsilon^2 D_n^4} = \frac{4c^2}{\varepsilon^2 D_n^2} \underset{n\to\infty}{\longrightarrow} 0\,,$$

weil $D_n^2 \underset{n\to\infty}{\longrightarrow} \infty$. □

Kapitel 6

Monte–Carlo–Simulation von Zufallsvariablen

Man kann ein Objekt der Wahrscheinlichkeitstheorie (z.B., ein Zufallselement) erst tief verstehen, wenn man es auf dem Rechner nachbilden (d.h., *simulieren*) kann. Der Ausdruck *Monte–Carlo– Simulation* wurde erstmals 1949 in der Arbeit [20] geprägt. Seine Urheber amerikanische Mathematiker *Nicholas Metropolis* und *Stanislaw M. Ulam*[1] wollten dabei verdeutlichen, dass Simulationsvorgänge dem Roulette–Spiel (wofür der Stadtteil Monte–Carlo im Fürstentum Monaco bekannt ist) in einer gewissen Hinsicht ähnlich sind. Die Basis solcher Simulationen bilden sogenannte *pseudozufällige Zahlen*, die als unabhängige Realisierungen einer auf $(0, 1)$ gleichverteilten Zufallsvariable interpretiert werden. Mitte des 20. Jahrhunderts wurden für die Erzeugung solcher Zahlenfolgen noch physikalische Geräte benutzt, die natürliche Fluktuationen (d.h., *Rauschen*) von z.B. elektrischer Spannung gemessen haben, und bei Überschreitung eines vorgegebenen Spannungsniveaus Eins, und sonst Null als binäre Zufallszahlen generiert haben. Diese binären Zahlenfolgen wurden als eine Binärdarstellung einer zufälligen Zahl auf dem Intervall $(0, 1)$ interpretiert. So generierte Zahlenfolgen wurden in Tabellen zufälliger Zahlen gespeichert. Seit der Entwicklung mächtiger Computer gelten solche direkten Methoden als antiquiert. Sie wurden durch rechnergestützte Generatoren von Pseudozufallszahlen ersetzt.

6.1 Pseudozufallszahlen

Es mag überraschend erscheinen, dass auf einem Rechner, der deterministisch arbeitet und so keine Zufälle zulässt, zufällige Zahlen simuliert werden

[1] Die Monte-Carlo Methoden wurden während des 2. Weltkrieges von Metropolis sowie Ulam zusammen mit den amerikanischen Physikern *Enrico Fermi* und *John von Neumann* in Los Alamos, USA im Rahmen des berüchtigten *Manhattan-Projekts* entwickelt. In dem Projekt ging es um die Erschaffung der ersten Atombombe.

können. In Wirklichkeit liefert der Rechner eine periodische Folge von deterministischen Zahlen, wobei die Periode jedoch ausreichend groß ist. Somit wiederholen sich die Werte der *pseudozufälligen* Zahlen selbst bei relativ langen Simulationsstudien nicht. Man sagt, dass eine Folge von pseudozufälligen Zahlen eine Zufallsvariable X *simuliert*, wenn diese Folge annähernd dieselben statistischen Eigenschaften aufweist wie eine Stichprobe von unabhängigen Realisierungen von X. Dies kann mit Hilfe statistischer Tests auf Gleichverteiltheit wie z.B. dem Kolmogorov–Smirnov–Test, χ^2–Test, usw. nachgewiesen werden.

Wie kann man eine $\mathcal{U}(0,1)$–verteilte Zufallsvariable simulieren? Auf dem Rechner wird diese absolut stetige Verteilung durch eine diskrete Gleichverteilung ersetzt, die diese ausreichend gut approximiert. Wenn wir z.B. pseudozufällige Zahlen mit einer Genauigkeit bis auf d Dezimalstellen brauchen, so kann der Wertebereich dieser diskreten Gleichverteilung als

$$\{k/10^d : k = 1, \ldots, 10^d - 1\}$$

gewählt werden. Jeder Wert $k/10^d$ wird mit Wahrscheinlichkeit

$$p_k = \frac{1}{10^d - 1}$$

angenommen.

Die Klasse der Methoden, die solche Verteilungen simulieren können, heißt *Generatoren pseudozufälliger Zahlen*. Sie unterscheiden sich in ihren Eigenschaften und in ihrer Komplexität. Die meisten modernen Generatoren arbeiten iterativ, also liefern eine Zahlenfolge $x_k = G(x_{k-1})$, $k \in \mathbb{N}$, wobei $G : (0,1) \to (0,1)$ eine Abbildung ist und der Anfangswert $u_0 \in (0,1)$ fixiert wird. Es ist klar, dass alle Punkte $(x_k, G(x_k)) \in [0,1]^2$ auf der Grafik von G liegen, somit soll G gewählt werden, sodass ihre Grafik möglichst dicht das ganze Einheitsquadrat $[0,1]^2$ füllt. Erst dann haben die Punkte $(x_k, G(x_k))$ eine Chance, pseudogleichverteilt auf $[0,1]^2$ auszusehen. Als ein natürlicher Kandidat für eine solche Funktion gilt $G(x) = \lfloor ax \rfloor$, $x \in (0,1)$, wobei $\lfloor ax \rfloor$ der ganze Teil von ax und $a > 0$ eine sehr große Zahl sind.

Hier werden wir zwei einfache Generatoren dieser Art kennen lernen. Weitere Generatoren können den Büchern [17, 21, 11, 29, 36, 13] entnommen werden.

1. *Residuenmethode*[2]: Seien $a, n \in \mathbb{N}$, wobei n und $(n-1)/2$ Primzahlen sind. Es gelte zusätzlich $a^{(n-1)/2} \equiv -1 \pmod{n}$. Definiere die Folge

$$u_k = a u_{k-1} \pmod{n}, \quad k \in \mathbb{N}, \tag{6.1}$$

wobei der Anfangswert $u_0 \in \{1, \ldots, n-1\}$ der *Keim* der Folge (engl. *seed*) heißt. Somit werden $x_k = u_k/n$ als unabhängige Realisierungen

[2]Sie heißt auf Englisch *residual method* oder *congruential method*. Ihre Idee wurde von D.H. Lehmer (1949) vorgeschlagen.

der Zufallsvariable $X \sim \mathcal{U}(0,1)$ interpretiert. Man kann beweisen, dass die Folge $\{u_k\}$ für beliebiges u_0 die Periode $n-1$ hat. Somit kann x_k höchstens $n-1$ Werte annehmen. Beispielsweise erfüllen $a = 1000$ und $n = 2001179$ die obigen Voraussetzungen. Alternativ kann man $a = 5^{17}$, $n = 2^{42}$ mit $u_0 = 1$ benutzen, selbst wenn a und n die obigen Voraussetzungen nicht erfüllen. Die Periode einer solchen Folge wird gleich 2^{40} sein.

Wie genau ist der Generator (6.1)? Sei U_0 eine auf $\{1,\dots,n-1\}$ gleichverteilte Zufallsvariable. Somit erzeugt die Relation (6.1) die Folge $U_k \equiv aU_{k-1} \pmod n$ von Zufallsvariablen, auf deren Basis eine neue Folge $X_k = U_k/n$ konstruiert wird, die mit $X \sim \mathcal{U}(0,1)$ verglichen werden soll. Man kann zeigen, dass alle X_k identisch verteilt sind mit Mittelwert

$$\mathbb{E}\, X_k = \mathrm{E}\, X = 1/2$$

und Varianz

$$\mathrm{Var}\ X_k = \frac{n-2}{12n} \to \frac{1}{12} = \mathrm{Var}\ X,$$

falls $n \to \infty$. Die Zufallsvariablen X_k sind offensichtlich nicht unabhängig voneinander. Man kann zeigen, dass der Korrelationskoeffizient

$$Corr(X_k, X_{k+1}) = \frac{Cov(X_k, X_{k+1})}{\sqrt{\mathrm{Var}\ X_k\, Var\, X_{k+1}}} \approx 1/a$$

und somit nicht Null ist.

2. Seien $a = 10^m + 1$ für $m \in \mathbb{N}$, $b \in \mathbb{N}$ nicht teilbar durch 2 oder 5, und sei $n = 10^d$ für $d \in \mathbb{N}$. Definiere die Folge

$$u_k = au_{k-1} + b \pmod n, \quad k \in \mathbb{N} \tag{6.2}$$

für einen Keimwert u_0. Die Periode der Folge (6.2) ist $n-1$. Setze $x_k = u_k/n$, $k \in \mathbb{Z}_+$.

Um bei unterschiedlichen Simulationsvorgängen verschiedene Folgen (6.1)–(6.2) zu bekommen, soll deren Keimwert u_0 so oft wie möglich geändert werden. Eine Ausnahme aus dieser Regel bilden Berechnungen, bei denen dieselbe Folge von Pseudozufallszahlen verwendet werden soll.

Im Folgenden nehmen wir an, dass wir ausreichend gute unabhängige Realisierungen der Zufallsvariablen $U \sim \mathcal{U}(0,1)$ generieren können. Um andere Zufallsvariablen auf Basis einer Gleichverteilung simulieren zu können, gibt es eine Reihe von Methoden wie z.B. die *Inversionsmethode* oder die *Akzeptanz– und Verwerfungsmethode.*

6.2 Inversionsmethode

Die Simulation von Zufallsvariablen X auf dem Rechner ist möglich, falls ihre Quantilfunktion F_X^{-1} explizit berechnet werden kann. Insbesondere ist es der Fall, wenn die Verteilungsfunktion F_X strikt monoton steigend ist, und damit die Quantilfunktion F_X^{-1} mit der herkömmlichen Inversen von F_X übereinstimmt. Zur Simulation erzeugt man dann eine *Pseudozufallszahl* u, die ein Zufallsgenerator des Rechners hergibt, und die (näherungsweise) einer Realisierung der Zufallsvariablen $U \sim \mathcal{U}(0,1)$ entspricht. Der Ausdruck $F_X^{-1}(u)$ liefert dann eine Realisierung von X.

Die Inversionsmethode gehört zur Klasse der sog. *Transformationsmethoden* zur Simulation von Zufallsvariablen. Diese bauen auf den Zusammenhängen der Art $X \stackrel{d}{=} T(Y)$ auf, wobei man den Zufallsvektor $Y = (Y_1, Y_2, \ldots, Y_m)$, $m \in \mathbb{N}$ simulieren kann, und die messbare Transformation $T : A \to \mathbb{R}$ findet, $A \in \mathcal{B}_{\mathbb{R}^m}$, so dass man aus Y die gesuchte Zufallsvariable X bekommt.

Beispiel 6.2.1

1. *Exponentialverteilung*: Sei $X \sim Exp(\lambda)$ für ein $\lambda > 0$. Ihre Quantilfunktion ist gegeben durch $F_X^{-1}(y) = -\lambda^{-1}\log(1-y)$, $y \in [0,1)$. Damit kann X per Inversionsmethode durch

$$X \stackrel{d}{=} -\lambda^{-1}\log(1-U) \stackrel{d}{=} -\lambda^{-1}\log U, \quad U \sim \mathcal{U}(0,1)$$

simuliert werden, weil hier $1 - U \stackrel{d}{=} U$ gilt. Da $P(U = 1) = P(U = 0) = 0$, sind Realisierungen von X f.s. endlich.

2. *Bernoulli–Verteilung*: Um $X \sim Ber(p)$, $p \in (0,1)$ zu simulieren, kann die Inversionsmethode wie folgt verwendet werden. Es gilt

$$F_X(x) = \begin{cases} 1, & x \geq 1, \\ 1-p, & x \in [0,1), \\ 0, & x < 0 \end{cases}$$

und somit $F_X^{-1}(y) = I(y > 1-p)$. Daher simuliere

$$X = I(U > 1-p) \stackrel{d}{=} I(1-U < p) \stackrel{d}{=} I(U \leq p)$$

für $U \sim \mathcal{U}(0,1)$. Die Simulation weiterer diskret verteilten Zufallsvariablen mittels Inversionsmethode wird in Abschnitt 6.5 besprochen.

3. *Cauchy-Verteilung*: Eine Zufallsvariable $X \sim Cauchy(0,1)$ kann mit Hilfe der Inversionsmethode wie

$$X \stackrel{d}{=} -\cot(\pi U), \quad U \sim \mathcal{U}(0,1)$$

simuliert werden, da die Quantilfunktion der Cauchy-Verteilung explizit bekannt ist. Alternativ kann eine Transformationsmethode verwendet werden, indem man folgendes Resultat nutzt. Es gilt, dass

$X \stackrel{d}{=} Y_1/Y_2$, wobei Y_1, Y_2 unabhängige $N(0,1)$–verteilte Zufallsvariablen sind, die z.B. mit Hilfe der Polar–Methode simuliert werden, vgl. Abschnitt 6.4. Hier handelt es also um die Transformation $T(y_1, y_2) = y_1/y_2$. Da der Wert $Y_2 = 0$ mit Wahrscheinlichkeit Null angenommen wird, ist der Wert von $T(Y_1, Y_2)$ f.s. endlich.

4. *Pareto–Verteilung*: Es lässt sich leicht zeigen, dass für $X \sim Par(\alpha, \mu)$, $\alpha, \mu > 0$ gilt $X \stackrel{d}{=} \mu U^{-1/\alpha}$, wobei $U \sim \mathcal{U}(0,1)$.

Da Quantilfunktionen nicht immer analytisch gegeben sind (wie etwa im Fall einer Normalverteilung), sind somit natürliche Grenzen für die Verwendung der *Inversionsmethode* zur Simulation von Zufallsvariablen gesetzt. Deshalb gibt es weitere Simulationsmethoden für Zufallsvariablen, welche auf anderen Ideen basieren, wie z.B. die *Akzeptanz– und Verwerfungsmethode*, die wir gleich beschreiben werden.

6.3 Akzeptanz– und Verwerfungsmethode

Es sei $X : \Omega \to \mathbb{R}$ eine Zufallsvariable auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, P)$, die absolut stetig verteilt ist mit Dichte f_X. Betrachten wir die Klasse der auf $\mathbb{R}$ integrierbaren Funktionen $f \geq 0$, die proportional zu f_X sind: $f(x) = cf_X(x)$, $x \in \mathbb{R}$ für ein $c > 0$. Somit gilt

$$f_X(x) = \frac{f(x)}{\int\limits_{\mathbb{R}} f(y)\, dy}, \quad x \in \mathbb{R}\,. \tag{6.3}$$

Nun wählen wir eine konkrete Funktion f aus dieser Klasse aus.

Nehmen wir an, dass eine auf $\mathbb{R}$ Lebesgue–integrierbare Funktion $g \geq 0$ mit den folgenden Eigenschaften existiert:

- $g(x) \geq f(x)$ für alle $x \in \mathbb{R}$,
- Wir können eine Zufallsvariable Y, die absolut stetig verteilt ist mit Dichte
$$g_Y(x) = \frac{g(x)}{\int\limits_{\mathbb{R}} g(y)\, dy}, \quad x \in \mathbb{R}\,, \tag{6.4}$$
simulieren.

Die *Akzeptanz– und Verwerfungsmethode* besteht aus den folgenden Schritten:

1. Simuliere die Zufallsvariablen Y und $U \sim \mathcal{U}(0,1)$ unabhängig voneinander.
2. Falls $Ug(Y) \leq f(Y)$, liefere Y.

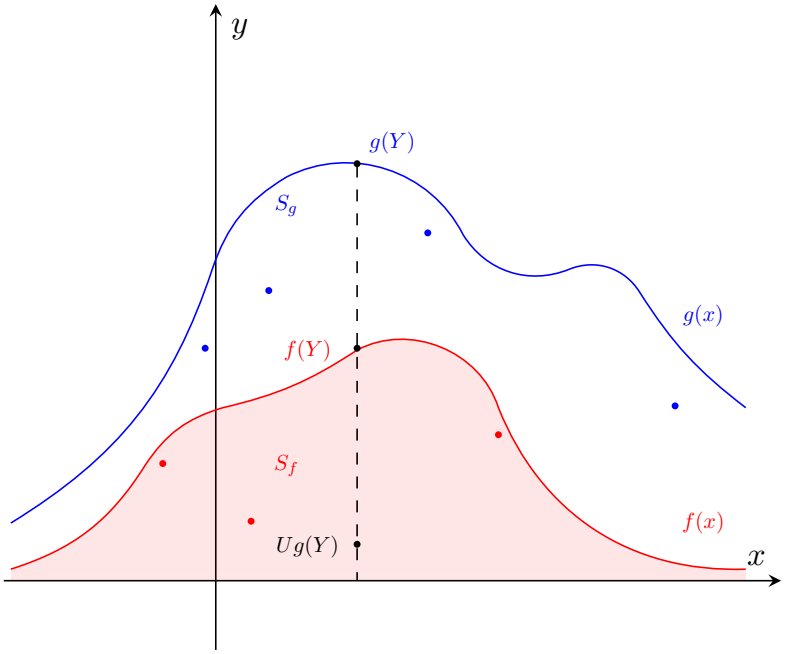


Abbildung 6.1: Akzeptanz– und Verwerfungsmethode.

3. Andernfalls, gehe zu 1.

Diese Schritte sind in Abbildung 6.1 veranschaulicht.

Im Folgenden bedeutet die Bezeichnung $Y \mid A$ die Einschränkung einer Zufallsvariablen $Y : \Omega \to \mathbb{R}$ auf das Ereignis $A \in \mathcal{F}$, oder, genauer gesagt, $Y(\omega)$ für $\omega \in A$.

Satz 6.3.1 Seien die Zufallsvariablen X und Y wie oben eingeführt. Die Zufallsvariable

$$\widetilde{X} = Y \mid \{Ug(Y) \leq f(Y)\} \tag{6.5}$$

hat dieselbe Verteilung wie X.

Bemerkung 6.3.1 Intuition Akzeptanz– und Verwerfungsmethode Der geometrische Sinn der obigen Vorgehensweise ist folgender: Falls ein zufälliger Punkt mit Koordinaten, die gleichverteilt auf S_g sind, unter den Graph von f fällt, wird seine x–Koordinate als Ergebnis des Algorithmus ausgegeben. Alle Punkte, die oberhalb des Graphes von f fallen, werden verworfen; vgl. Abbildung 6.1.

In jedem Schritt der Akzeptanz– und Verwerfungsmethode wird der Vorschlag Y akzeptiert, falls $Ug(Y) \leq f(Y)$. Sonst wird ein neuer unabhängiger Wert von Y generiert. Sei M die Anzahl solcher Schritte bis zur ersten Akzeptanz von Y. Die Größe M ist offensichtlich geometrisch verteilt mit Parameter

$$p = P(Ug(Y) \leq f(Y)).$$

Somit ist die Erfolgswahrscheinlichkeit gleich

$$p = \int_{\mathbb{R}} P\left(U \le \frac{f(y)}{g(y)}\right) g_Y(y)\, dy = \int_{\mathbb{R}} \frac{f(y)}{g(y)} \frac{g(y)}{|S_g|}\, dy = \frac{|S_f|}{|S_g|}\,.$$

Folglich ist die mittlere Anzahl der notwendigen Simulationsschritte gleich

$$\mathbb{E}\, M = \frac{1}{p} = \frac{|S_g|}{|S_f|} = \frac{\int_{\mathbb{R}} g(x)\, dx}{\int_{\mathbb{R}} f(x)\, dx} > 1. \tag{6.6}$$

Das heißt, je besser die obere Schranke g für f ist, desto schneller ist im Mittel die Simulation. Die Varianz von M ist gleich

$$\text{Var}\ \ M = \frac{1}{p^2} - \frac{1}{p} = \frac{1}{p}\left(\frac{1}{p} - 1\right) = \frac{|S_g|}{|S_f|}\left(\frac{|S_g|}{|S_f|} - 1\right).$$

Ein Beispiel von Verwendung der Akzeptanz– und Verwerfungsmethode zur Simulation der Normalverteilung geben wir in Abschnitt 6.4.1.

Übungsaufgabe 6.3.1 (Gamma–Verteilung)
Wie lautet der Akzeptanz– und Verwerfungsalgorithmus zur Simulation von $X \sim \Gamma(1, a)$, $a > 0$, $a \neq 1$? Verwenden Sie dabei die Funktion

$$g(x) = \begin{cases} x^{a-1}\, I_{(0 \le x < 1)} + e^{-x}\, I_{(x \ge 1)}, & a < 1, \\ \frac{x^{\lambda-1}}{(x^\lambda + a^\lambda)^2} I_{(x \ge 0)}, & a > 1, \end{cases}$$

mit $\lambda = \sqrt{2a-1}$. Geben Sie die mittlere Anzahl und die Varianz der notwendigen Simulationsschritte an.

6.4 Simulation der Normalverteilung

Wie bereits erwähnt wurde, gibt es keinen geschlossenen Ausdruck für die Quantilfunktion der Standardnormalverteilung, was die Verwendung der Inversionsmethode erschwert. Deshalb werden in diesem Abschnitt andere Simulationsmöglichkeiten für $N(0,1)$ aufgezeigt, z.B. mit Hilfe der Akzeptanz– und Verwerfungsmethode oder der Box–Muller–Transformation. Diese Algorithmen, im Gegensatz zu den Approximationsansätzen basierend auf dem zentralen Grenzwertsatz (Theorem 5.2.1), sind genau, also liefern exakte Realisierungen von $N(0,1)$–verteilter Zufallsvariable. Eine Realisierung von einer Zufallsvariablen $Z \sim N(\mu, \sigma^2)$ bekommt man durch die Relation $Z = \mu + \sigma X$, wobei X eine Realisierung von $N(0,1)$ darstellt.

6.4.1 Akzeptanz– und Verwerfungsmethode für $N(0,1)$

Sei die Zufallsvariable X standardnormalvertteilt mit Dichte

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \mathbb{R}.$$

Es gilt $f_X(x) \leq g(x) = \sqrt{\frac{e}{2\pi}} e^{-|x|}$, $x \in \mathbb{R}$, weil aus

$$x^2 - 2|x| + 1 = (x \pm 1)^2 \geq 0$$

die Ungleichung $e^{-x^2/2} \leq e^{1/2-|x|}$ folgt. Sei die Zufallsvariable Y absolut stetig verteilt mit einer Dichte, die proportional zu $g(x)$ ist. Dies bedeutet, dass $|Y| \sim Exp(1)$ und daher $|Y| \stackrel{d}{=} -\log V$ für eine Zufallsvariable $V \sim \mathcal{U}(0,1)$; vgl. Beispiel 6.2.1 1. Für ein $U \sim \mathcal{U}(0,1)$ sieht das Akzeptanz–Kriterium $Ug(Y) \leq f_X(Y)$ folgendermaßen aus:

$$U \frac{1}{\sqrt{2\pi}} e^{1/2-|Y|} \leq \frac{1}{\sqrt{2\pi}} e^{-|Y|^2/2},$$

was sich umschreiben lässt wie

$$U e^{1/2-(-\log V)} \leq e^{-(\log^2 V)/2},$$

wobei die Zufallsvariablen U und V stochastisch unabhängig sind. Beide Seiten der Ungleichung logarithmierend, bekommt man

$$\log U + \log V + \frac{1}{2} \leq -\frac{1}{2} \log^2 V,$$

oder, äquivalent dazu,

$$2 \log U \leq -(\log V + 1)^2. \tag{6.7}$$

Falls die Bedingung (6.7) erfüllt ist, wird der Wert von $-\log V$ als eine Realisierung von $|X|$ akzeptiert. Da die Zufallsvariable $|X|$ symmetrisch ist, kann das Vorzeichen von X stochastisch unabhängig von U und V mit Wahrscheinlichkeit $1/2$ gewählt werden. Mit anderen Worten, es gilt $X \stackrel{d}{=} (2B-1) \log V$, wobei $B \sim Ber(1/2)$ eine von U, V stochastisch unabhängige Zufallsvariable ist. Offensichtlich nimmt $2B-1$ Werte ± 1 mit Wahrscheinlichkeit $1/2$ an. Die Simulation von B als $I(2W \leq 1)$ für $W \sim \mathcal{U}(0,1)$ ist in Beispiel 6.2.1 2 bereits beschrieben worden.

Zusammengefasst, sieht der Algorithmus zur Simulation von $X \sim N(0,1)$ wie folgt aus:

Bemerkung 6.4.1 Simulation von $N(0,1)$ mit Akzeptanz– und Verwerfungsmethode

1. Simuliere $U, V, W \sim \mathcal{U}(0,1)$ stochastisch unabhängig voneinander.

2. Falls $2\log U \leq -(\log V + 1)^2$ gilt, liefere $X = (2I(2W \leq 1) - 1)\log V$.

3. Sonst gehe zu Schritt 1.

Dieser Algorithmus ist ziemlich schnell, denn die mittlere Anzahl M der Simulationsschritte ist nach (6.6) gleich

$$\mathbb{E}\, M = \frac{|S_g|}{|S_{f_X}|} = \frac{\sqrt{\frac{e}{2\pi}} \int\limits_{\mathbb{R}} e^{-|x|}\, dx}{1} = \sqrt{\frac{2e}{\pi}} \int\limits_0^\infty e^{-x}\, dx = \sqrt{\frac{2e}{\pi}} \approx 1.315$$

mit der Varianz Var $M \approx 1.315(1.315 - 1) = 0.414225$.

6.4.2 Box–Muller-Transformation

Eine Alternativmethode zur Simulation von der Standardnormalverteilung, auch *Polarmethode* genannt (vgl. z.B. [29, Sektion 5.3]), liefert die sog. *Box–Muller–Transformation*

$$X \stackrel{d}{=} \sqrt{-2\log U}\cos(2\pi V) \stackrel{d}{=} \sqrt{-2\log U}\sin(2\pi V),$$

wobei $U, V \sim \mathcal{U}(0,1)$ stochastisch unabhängig sind. Sie basiert auf folgendem Lemma:

Lemma 6.4.1

1. Seien (R, Θ) die Polarkoordinaten des Zufallsvektors (X_1, X_2), wobei X_1 und X_2 zwei $N(0,1)$–verteilte, stochastisch unabhängige Zufallsvariablen sind. Dann sind R und Θ stochastisch unabhängige Zufallsvariablen, und es gilt $R^2 \sim Exp(1/2)$, $\Theta \sim \mathcal{U}(0, 2\pi)$.

2. Falls $R^2 \sim Exp(1/2)$ und $\Theta \sim \mathcal{U}(0, 2\pi)$ stochastisch unabhängige Zufallsvariablen sind, dann sind die Zufallsvariablen $X_1 = R\cos\Theta$ und $X_2 = R\sin\Theta$ standardnormalverteilt und stochastisch unabhängig.

Bemerkung 6.4.2 (Rayleigh–Verteilung) Die oben eingeführte Zufallsvariable $R = \sqrt{X_1^2 + X_2^2}$ ist *Rayleigh–verteilt* mit Parameter $\sigma = 1$. Generell wird die Dichte der *Rayleigh–Verteilung mit Parameter* $\sigma > 0$ (wir schreiben abkürzend $RL(\sigma)$) durch den Ausdruck

$$f(x) = \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}} I(x \geq 0)$$

definiert. Dadurch bekommt die Tailfunktion von $R_\sigma \sim RL(\sigma)$ folgende simple Form:

$$\bar{F}_{R_\sigma}(x) = I(x < 0) + e^{-\frac{x^2}{2\sigma^2}} I(x \geq 0).$$

Die $RL(\sigma)$-Verteilung ist nach dem britischen Physiker und Nobel–Preisträger *John W. Strutt, 3. Baron Rayleigh* (1842–1919) benannt. Sie dient beispielsweise zur Modellierung von 10-minutigen Mittelwerten der Windgeschwindigkeiten.

Per Definition der χ^2–Verteilung gilt

$$R^2 = X_1^2 + X_2^2 \sim \chi_2^2 = \Gamma(1/2, 1) = Exp(1/2).$$

Außerdem ist die Relation $R_\sigma \stackrel{d}{=} \sigma R$ für jedes $\sigma > 0$ offensichtlich, insbesondere $R_1 \stackrel{d}{=} R$. Damit gilt

$$R_\sigma \stackrel{d}{=} \sqrt{Z_1^2 + Z_2^2},$$

wobei $Z_1, Z_2 \sim N(0, \sigma^2)$ stochastisch unabhängige Zufallsvariablen sind.

Nach Lemma 6.4.1 2 kann $X \sim N(0,1)$ als $R\cos\Theta$ oder $R\sin\Theta$ simuliert werden, wobei $R \stackrel{d}{=} \sqrt{-2\log U}$ und $\Theta \stackrel{d}{=} 2\pi V$ für $U, V \sim \mathcal{U}(0,1)$. Dadurch erhalten wir folgenden direkten (d.h., nicht–iterativen) Algorithmus zur Simulation von $N(0,1)$:

Bemerkung 6.4.3 Simulation von $N(0,1)$ mit Polarmethode

1. Simuliere stochastisch unabhängige Zufallsvariablen $U, V \sim \mathcal{U}(0,1)$.
2. Liefere $X = \sqrt{-2\log U}\cos(2\pi V)$ oder $X = \sqrt{-2\log U}\sin(2\pi V)$.

Die Zufallsvariable $Z \sim N(\mu, \sigma^2)$ kann als $\mu + \sigma X$ für beliebige $\mu \in \mathbb{R}$, $\sigma > 0$ simuliert werden.

6.5 Simulation von diskret verteilten Zufallsvariablen

Sei X eine diskret verteilte Zufallsvariable, die Werte $x_k \in \mathbb{R}$, $k \in \mathbb{N}$ mit Wahrscheinlichkeiten $p_k = P(X = x_k)$ annimmt. Es gilt $\sum_{k=1}^{\infty} p_k = 1$. Sei $P_k = p_1 + \ldots + p_k$ für alle k. Wie simuliert man X?

In Beispiel 6.2.1 2 wurde eine Bernoulli-verteilte Zufallsvariable per Inversionsmethode (vgl. Abschnitt 6.2) simuliert. Diese Methode kann auch zur Simulation von allgemeinen diskret verteilten Zufallsvariablen X wie folgt benutzt werden:

Bemerkung 6.5.1 Simulation von diskreten Verteilungen mit Inversionsmethode

1. Simuliere eine Zufallsvariable $U \sim \mathcal{U}(0,1)$.

2. Liefere

$$\widetilde{X} = \begin{cases} x_1, & U < P_1, \\ x_2, & P_1 \le U < P_2, \\ \vdots, & \vdots \\ x_k, & P_{k-1} \le U < P_k, \\ \vdots & \vdots \end{cases} \tag{6.8}$$

als eine Realiserung von X.

Es gilt offensichtlich $\widetilde{X} \stackrel{d}{=} X$.

Diese Methode kann effizient nur für relativ kleine endliche Wertebereiche $C = \{x_1, \ldots, x_n\}$ von X (d.h., falls $p_k = 0$ für $k > n$) angewandt werden, weil für große n zu viele Fälle unterschieden werden müssen. Als Ausweg aus dieser Situation können *Markov–Ketten–Monte–Carlo–Methoden* genannt werden, die wegen ihrer Komplexität nicht in diesem Einführungstext behandelt werden; siehe z.B. Bücher [1, 17, 11, 13, 30].

Beispiel 6.5.1 (Binomialverteilung)
Um eine Zufallsvariable $X \sim Bin(n, p)$ zu simulieren, kann die Darstellung

$$X \stackrel{d}{=} X_1 + \ldots + X_n$$

benutzt werden, wobei X_k, $k = 1, \ldots, n$ stochastisch unabhängige $Ber(p)$–verteilte Zufallsvariablen sind. Aus Beispiel 6.2.1 2 folgt, dass

$$X \stackrel{d}{=} I(U_1 \le p) + \ldots + I(U_n \le p)$$

für stochastisch unabhängige Zufallsvariablen $U_1, \ldots, U_n \sim \mathcal{U}(0, 1)$. Allerdings ist dieser Ansatz für große n offensichtlich ineffizient. In diesem Fall kann die Methode (6.8) benutzt werden, wobei die Wahrscheinlichkeiten $p_k = \binom{n}{k} p^k (1-p)^{n-k}$, $k = 0, \ldots, n$ rekursiv durch die Formel

$$p_k = \frac{n-k+1}{k} \frac{p}{1-p} \cdot p_{k-1}, \quad k = 1, \ldots, n, \qquad p_0 = (1-p)^n$$

berechnet werden, die einen Spezialfall der sog. *Panjer–Rekursion* darstellt, vgl. z.B. [24, 35], [28, Theorem 4.3.1]. Für kleine $\mathbb{E}\, X = np$ werden die Vergleiche in (6.8) in natürlicher Reihenfolge durchgeführt. Für große np ist es effizienter mit $P_{\lfloor np \rfloor}$ anzufangen. Falls n groß und $p \ll 0.25$ klein ist, kann man von der Poisson–Approximation in Theorem 3.3.1 Gebrauch machen, und zwar gilt $X \approx \widetilde{X} \sim Poisson(np)$, siehe Beispiel 6.5.2 2. Hier ist allerdings mit einem Approximationsfehler zu rechnen.

Alternativ kann die Akzeptanz– und Verwerfungsmethode verwendet werden. Sei hierfür $X : \Omega \to C$ eine diskret verteilte Zufallsvariable mit Zähldichte $\{p_k\}_{k\in\mathbb{N}}$ und (evtl. unendlichem) Wertebereich $C = \{x_1, \ldots, x_n, \ldots\}$. Für eine stetig verteilte Zufallsvariable $Z : \Omega \to \mathbb{R}_+$ mit der stückweise konstanten Dichte

$$f_Z(x) = p_{\lfloor x \rfloor} I(x \geq 1)$$

gilt offensichtlich die Gleichung $X \stackrel{d}{=} x_{\lfloor Z \rfloor}$. Die Zufallsvariable Z kann somit behilflich sein, die Nummer $k \in \mathbb{N}$ des Zustandes $x_k \in C$ von X zu simulieren, was zusammenfassend folgendermassen festzuhalten ist:

Bemerkung 6.5.2 Simulation von diskreten Verteilungen mit Akzeptanz– und Verwerfungsmethode

1. Simuliere die Zufallsvariable Z mit Hilfe der Akzeptanz– und Verwerfungsmethode aus Abschnitt 6.3.
2. Liefere $x_{\lfloor Z \rfloor}$ als eine Realisierung von X.

Es gibt eine Reihe von *ad hoc* Algorithmen für die Simulation von diskret verteilten Zufallsvariablen X, die auf speziellen Eigenschaften der Verteilung von X beruhen. Manche von ihnen werden hier exemplarisch aufgeführt.

Beispiel 6.5.2 (Ad hoc Methoden)

1. *Geometrische Verteilung*:
   Für die Zufallsvariable $X \sim Geo(p)$, $0 < p < 1$, gilt die Gleichung

   $$X \stackrel{d}{=} \lfloor \log_{1-p} U \rfloor + 1, \qquad U \sim \mathcal{U}(0,1),$$

   die eine geeignete Transformation von Pseudozufallszahl U zur Simulation von X liefert. Um diese Darstellung zu beweisen, schreibt man

   $$\begin{aligned} P(\lfloor \log_{1-p} U \rfloor + 1 = k) &= P(k - 1 \leq \log_{1-p} U < k) \\ &= P\left((1-p)^k < U \leq (1-p)^{k-1}\right) \\ &= (1-p)^{k-1} - (1-p)^k = p(1-p)^{k-1}, \quad k \in \mathbb{N}. \end{aligned}$$

2. *Poisson–Verteilung*:
   Sei $X \sim Poisson(\lambda)$. Es gilt die Darstellung

   $$X = inf\{k \in \mathbb{Z}_+ : \sum_{j=1}^{k+1} Y_j > \lambda\} \tag{6.9}$$

   für stochastisch unabhängige Zufallsvariablen $Y_j \sim Exp(1)$. Die Inversionsmethode ergibt $Y_j \stackrel{d}{=} -\log U_j$ für $U_j \sim \mathcal{U}(0,1)$, $j \in \mathbb{N}$. Die

Bedingung in (6.9) kann somit als $-\sum_{j=1}^{k+1} \log U_j > \lambda$ oder $\prod_{j=1}^{k+1} U_j < e^{-\lambda}$ umgeschrieben werden. Fassen wir diese Überlegungen in einem Algorithmus zusammen:

(a) Setze $k = 0$, $T = 1$.

(b) Simuliere $U \sim \mathcal{U}(0,1)$ und bilde $T = UT$.

(c) Falls $T \geq e^{-\lambda}$, setze $k = k + 1$ und gehe zurück zum zweiten Schritt.

(d) Andernfalls, liefere k als Realisierung von X.

Da $\mathbb{E}\, X = \lambda$, ist die mittlere Anzahl M der Simulationsschritte hier gleich $\lambda + 1$. Somit wird dieser Algorithmus für große λ nicht mehr effizient.

Für große λ kann man die Inversionsmethode (6.8) anwenden, wobei die Zähldichte $p_k = e^{-\lambda}\frac{\lambda^k}{k!}$, $k \in \mathbb{Z}_+$ durch die Panjer–Rekursion

$$p_k = \frac{\lambda}{k} p_{k-1}, \quad k \in \mathbb{N}, \qquad p_0 = e^{-\lambda}$$

schnell berechnet werden können, vgl. z.B. [28, Theorem 4.3.1]. Um die Suche in (6.8) für große λ zu optimieren, wird wegen $E\, N = \lambda$ zunächst U mit $P_{\lfloor\lambda\rfloor}$ verglichen. Falls $U < P_{\lfloor\lambda\rfloor}$ wird weiterhin geprüft, ob $U < P_{\lfloor\lambda\rfloor - 1}$ gilt, usw. Man setzt dann $X = \min\{k : U < P_k\}$. Analog geht man im Falle $U \geq P_{\lfloor\lambda\rfloor}$ vor.

Lemma 6.5.1 Für große λ, die mittlere Anzahl der Vergleiche ist dabei ungefähr gleich $1 + 0.798\sqrt{\lambda}$.

Nach dem zentralen Grenzwertsatz gilt

$$\frac{X - \lambda}{\sqrt{\lambda}} \xrightarrow{d} Y \sim N(0,1), \quad \lambda \to +\infty. \tag{6.10}$$

Hieraus folgt die Relation $X \approx \lambda + \sqrt{\lambda} Y$, die für $\lambda \geq 100$ als Approximation

$$X \approx \lfloor \lambda + 0.5 + \sqrt{\lambda} Y \rfloor, \quad Y \sim N(0,1)$$

nutzbar ist. Es gibt auch genauere approximative Simulationsmethoden für die Poisson–Verteilung wie etwa die *Anscombe–* und *Peizer und Pratt–Approximationen*; vgl. [7, S. 35–37, 142].

6.6 Markov-Ketten

Markov-Ketten (nach Andrei Andrejewitsch Markov, 1856-1922) bilden eine grundlegende Klasse stochastischer Modelle für Folgen von Zufallsvariablen

$X_0, X_1, \ldots$, die nicht unabhängig sind und daher eine gewisse *Abhängigkeitsstruktur* aufweisen. Sie können die zeitliche Entwicklung von Objekten, Sachverhalten, Systemen etc. in diskreten Zeitschritten $t = 0, 1, \ldots$ beschreiben, wobei zu jedem Zeitpunkt jeweils nur eine von endlich vielen Ausprägungen angenommen werden kann.

Markov-Ketten finden unter anderem zahlreiche Anwendungen bei der mathematischen Analyse und Modellierung in der Finanz- und Versicherungsmathematik, den Lebenswissenschaften, der Ökonomie, der Warteschlangentheorie und der Informatik. Außerdem werden Markov-Ketten zur Generierung von Pseudozufallszahlen und Pseudozufallsvektoren verwendet. Denn oftmals sind die interessierenden Fragestellungen und damit auch die entsprechenden stochastischen Modelle so komplex, dass ihre Verteilungen nicht mit den erläuterten (direkten) Algorithmen der vorherigen Abschnitte simuliert werden können. In einem solchen Fall ist es jedoch oft möglich, eine Markov-Kette zu konstruieren, deren asymptotische (Grenz-) Verteilung mit der zu simulierenden Verteilung übereinstimmt. Diese Art von Simulationsalgorithmen wird Markov Chain Monte Carlo (MCMC) genannt und in Abschnitt 6.7 erläutert.

6.6.1 Modellbeschreibung und Beispiele

Das stochastische Modell der Markov-Kette besteht aus drei Komponenten: dem (endlichen) Zustandsraum, der Anfangsverteilung und der Übergangsmatrix.

Zustandsraum Den Ausgangspunkt bildet die endliche Menge aller möglichen Zustände, die *Zustandsraum* der Markov-Kette genannt wird. Dabei nehmen wir an, dass jedes Element X_i der Markov-Kette $X_0, X_1, \ldots$ je $\ell \in \mathbb{N}$ verschiedene Zustände annehmen kann und identifizieren den Zustandsraum in der Regel mit der Menge $E = \{1, 2, \ldots, \ell\}$. Dies bedeutet *nicht*, dass die Zustände der Markov-Kette natürliche Zahlen sein müssen. Im Gegenteil, die möglichen ℓ Zustände können eine fast beliebig komplizierte Struktur aufweisen (z.B. reelle Zahlen, Vektoren, Matrizen, digitale Landkarten, mikroskopische 3D-Bilder, soziale Netzwerke, usw.). Jeder dieser (endlich vielen) möglichen Zustände wird dabei durch eine natürliche Zahl identifiziert.

Anfangsverteilung Für jedes $i \in E$ sei α_i die Wahrscheinlichkeit, dass sich das betrachtete Objekt, Sachverhalt bzw. System zum „Zeitpunkt" $n = 0$ im Zustand i befindet, d.h. $\alpha_i = P(X_0 = i)$. Das bedeutet, dass der Vektor $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_\ell)^\top$ eine Wahrscheinlichkeitsfunktion auf E bildet und damit $\alpha_i \in [0, 1]$ für jedes $i \in E$ und $\sum_{i=1}^{\ell} \alpha_i = 1$. Die Wahrscheinlichkeitsfunktion $\boldsymbol{\alpha}$ wird *Anfangsverteilung* der Markov-Kette genannt.

Übergangsmatrix Für jedes Paar $i, j \in E$ betrachten wir die (bedingte) Wahrscheinlichkeit p_{ij}, dass das betrachtete Objekt, Sachverhalt bzw. System in einem (Zeit-) Schritt aus dem Zustand i in den Zustand j übergeht. Das bedeutet, dass $p_{ij} \in [0, 1]$ für jedes $i, j \in E$ und $\sum_{j=1}^{\ell} p_{ij} = 1$ für jedes $i \in E$. Die $\ell \times \ell$ Matrix $\boldsymbol{P} = (p_{ij})_{i,j=1,\ldots,\ell}$ heißt *Übergangsmatrix* der Markov-Kette. Eine Matrix, die diese Voraussetzungen erfüllt, wird auch *stochastische Matrix* genannt.

Für jeden Zustandsraum $E = \{1, 2, \ldots, \ell\}$, jede Anfangsverteilung $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_\ell)^\top$ und jede Übergangsmatrix $\boldsymbol{P} = (p_{ij})$ kann nun der Begriff der zugehörigen Markov-Kette wie folgt eingeführt werden.

Definition 6.6.1 (Markov-Kette) Es sei $X_0, X_1, \ldots : \Omega \to E$ eine Folge von diskreten Zufallsvariablen mit Werten in der Menge $E = \{1, 2, \ldots, \ell\}$. Dann heißt $X_0, X_1, \ldots$ *Markov-Kette* mit Zustandsraum E, Anfangsverteilung $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_\ell)^\top$ und Übergangsmatrix $\boldsymbol{P} = (p_{ij})$, wenn

$$P(X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) = \alpha_{i_0} p_{i_0 i_1} \ldots p_{i_{n-1} i_n}$$

für beliebige $n = 0, 1, \ldots$ und $i_0, i_1, \ldots, i_n \in E$ gilt.

Bemerkung 6.6.1

1. Wenn die Folge der Zufallsvariablen $X_0, X_1, \ldots$ eine Markov-Kette bildet, dann ist diese eindeutig durch den Zustandsraum E, die Anfangsverteilung $\boldsymbol{\alpha}$ und die Übergangsmatrix $\boldsymbol{P}$ charakterisiert.

2. Die einzelnen Elemente $X_0, X_1, \ldots$ einer Markov-Kette sind im Allgemeinen weder unabhängig noch identisch verteilt.

3. Es existiert die folgende äquivalente Definitionsmöglichkeit von Markov-Ketten, die in der Literatur *Markov-Eigenschaft* genannt wird:
   Eine Folge $X_0, X_1, \ldots : \Omega \to E$ von diskreten Zufallsvariablen mit Werten in der Menge $E = \{1, 2, \ldots, \ell\}$ ist genau dann eine Markov-Kette mit Zustandsraum E und Übergangsmatrix $\boldsymbol{P} = (p_{ij})$, wenn

   $$\begin{aligned} &P(X_n = i_n \mid X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) \\ &\quad = P(X_n = i_n \mid X_{n-1} = i_{n-1}) = p_{i_{n-1} i_n}\,, \end{aligned}$$

   für beliebige $n = 1, 2, \ldots$ und $i_0, i_1, \ldots, i_n \in E$ mit

   $$P(X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) > 0.$$

   Diese Darstellung bedeutet, dass die bedingte Verteilung des (zufälligen) Zustandes X_n der Markov-Kette zum Zeitpunkt n vollständig durch den Zustand $X_{n-1} = i_{n-1}$ zum vorhergehenden Zeitpunkt $n-1$ bestimmt wird. Es ist also unwichtig, in welchem Zustand sich die Markov-Kette zu den Zeitpunkten $n-2, n-3, \ldots$ befunden hat.

Beispiel 6.6.1 1. *Produktivität eines Mitarbeiters*

Ein Mitarbeiter zieht am Ende jedes Arbeitstages Bilanz, ob er alle Aufgaben, die er sich vorgenommen hat, bearbeiten konnte oder nicht. Dazu betrachtet er den Zustandsraum $E = \{1, 2\}$, wobei 1 für „nicht alle Aufgaben bearbeitet" und 2 für „alle Aufgaben bearbeitet" steht. Es sei $X_0 = 2$ und für $n = 1, 2, \ldots$ gebe die Zufallsvariable $X_n : \Omega \to E$ an, ob der Mitarbeiter am n-ten Folgetag alle Aufgaben bearbeitet hat oder nicht. Da nicht erledigte Aufgaben am Folgetag bearbeitet werden müssen, sind die Zufallsvariablen $X_0, X_1, X_2, \ldots$ offensichtlich nicht unabhängig.

Es ist allerdings bekannt, dass der Mitarbeiter all seine Aufgaben an einem Tag mit Wahrscheinlichkeit 0,9 erledigt, wenn er schon am Vortag alle Aufgaben erledigen konnte. War das nicht der Fall, so sinkt diese Wahrscheinlichkeit auf 0,5. Die Folge $X_0, X_1, \ldots$ bildet dann eine Markov-Kette mit Zustandsraum E, Anfangsverteilung $\boldsymbol{\alpha} = (0, 1)^\top$ und Übergangsmatrix

$$\boldsymbol{P} = \begin{pmatrix} 0,5 & 0,5 \\ 0,1 & 0,9 \end{pmatrix}.$$

Ein allgemeinerer Fall liegt vor, wenn die konkreten Übergangswahrscheinlichkeiten nicht bekannt sind. Es sei $p \in (0, 1)$ die Wahrscheinlichkeit, dass der Mitarbeiter seine Aufgaben erledigen kann, wenn er am Vortag nicht alle Aufgaben erledigen konnte und $p' \in (0, 1)$ die Wahrscheinlichkeit, dass der Mitarbeiter nicht alle Aufgaben schafft, wenn er am Vortag alles erledigen konnte. Dann ist die Übergangsmatrix $\boldsymbol{P}$ wie folgt gegeben:

$$\boldsymbol{P} = \begin{pmatrix} 1-p & p \\ p' & 1-p' \end{pmatrix}.$$

2. *Diffusionsmodell nach Ehrenfest*

Das folgende Modell zur Beschreibung von Diffusionsvorgängen durch eine Membran wurde im Jahre 1907 von den Physikern Tatjana Ehrenfest-Afanassjewa (1876-1964) und Paul Ehrenfest (1880-1933) vorgeschlagen. Dabei geht es um die Modellierung des Wärmeaustausches zwischen zwei Systemen mit unterschiedlichen Temperaturen.

Wir betrachten ℓ Partikel, die auf zwei miteinander durchlässig verbundene, aber nach außen isolierte Behälter A und B verteilt sind. Angenommen zum Zeitpunkt $n-1$ befinden sich i Partikel in A. Dann wird eines der ℓ insgesamt vorhandenen Partikel rein zufällig ausgewählt und in den jeweils anderen Behälter überführt. Für die Anzahl

X_n von Partikeln in Behälter A zum Zeitpunkt n gilt somit entweder $X_n = i - 1$ mit Wahrscheinlichkeit $\frac{i}{\ell}$ (falls das ausgewählte Partikel im Behälter A war) oder $X_n = i + 1$ mit Wahrscheinlichkeit $\frac{\ell - i}{\ell}$ (falls das ausgewählte Partikel im Behälter B war). Die bedingte Verteilung von X_n wird also vollständig durch die Anzahl $X_{n-1} = i$ von Partikeln in A zum Zeitpunkt $n-1$ bestimmt. Des Weiteren beschreibe X_0 die zufällige Anzahl von Partikeln in A zum Zeitpunkt $n = 0$, deren Verteilung beliebig gewählt werden kann.

Die Zufallsvariablen $X_0, X_1, \ldots$ bilden dann eine Markov-Kette mit Zustandsraum $E = \{0, 1, \ldots, \ell\}$, beliebig zu wählender Anfangsverteilung $\boldsymbol{\alpha}$ und Übergangsmatrix $\boldsymbol{P} = (p_{ij})$, die wie folgt gegeben ist:

$$p_{ij} = \begin{cases} \dfrac{\ell - i}{\ell}\,, & \text{falls } i < \ell \text{ und } j = i + 1, \\ \dfrac{i}{\ell}\,, & \text{falls } i > 0 \text{ und } j = i - 1, \\ 0\,, & \text{sonst.} \end{cases}$$

Das Ehrenfest-Modell lässt sich auf viele weitere Anwendungen übertragen. Sei zum Beispiel ein soziales Netzwerk gegeben, dessen Knoten sich aufgrund ihrer Verknüpfungen in 2 Cluster A und B einteilen lassen. In regelmäßigen Abständen verändern sich die Kanten des Netzwerks jedoch, sodass ein zufällig ausgewählter Knoten nun von dem einem zum anderen Cluster wechselt. Wenn dabei die Wahrscheinlichkeit, dass ein Knoten in Cluster A ausgewählt wird, nur von der aktuellen Anzahl von Knoten in A abhängt, dann kann das Ehrenfest-Modell benutzt werden, um die Größe des Clusters A zu modellieren.

6.6.2 Rekursive Darstellung und n-Schritt Übergangswahrscheinlichkeiten

Um nachzuweisen, dass eine Folge $X_0, X_1, \ldots$ von Zufallsvariablen eine Markov-Kette bildet, muss die Gültigkeit der Bedingung in Definition 5.2 oder der dazu äquivalenten Markov-Eigenschaft gezeigt werden. Dies ist allerdings in vielen Fällen nicht explizit möglich. Alternativ kann man jedoch zeigen, dass eine Folge $X_0, X_1, \ldots$ eine Markov-Kette bildet, wenn sie eine bestimmte rekursive Darstellung mit unabhängigen und identisch verteilten Innovationen besitzt.

Satz 6.6.1 Es sei $E = \{1, \ldots, \ell\}$ eine beliebige endliche Menge und $(D, \mathcal{D})$ ein Messraum, wobei $D \subset \mathbb{R}$ eine nichtleere Menge und $\mathcal{D}$ eine σ-Algebra über D ist. Es sei weiterhin $Z_1, Z_2, \ldots : \Omega \to D$ eine Folge von unabhängigen und identisch verteilten Zufallsvariablen mit Werten in D („Innovationen" genannt) und $X_0 : \Omega \to E$ eine Zufallsvariable mit Werten in E, die unabhängig von $Z_1, Z_2, \ldots$ ist.

Schließlich sei $\varphi : E \times D \to E$ eine Funktion mit Werten in E, sodass $\{x \in D : \varphi(i,x) = j\} \in \mathcal{D}$ für beliebige $i, j \in E$. Dann bildet die Folge von Zufallsvariablen $X_0, X_1, \ldots : \Omega \to E$ mit $X_n = \varphi(X_{n-1}, Z_n)$ für $n \geq 1$ eine Markov-Kette mit Zustandsraum E, Anfangsverteilung $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_\ell)^\top$ und Übergangsmatrix $\boldsymbol{P} = (p_{ij})_{i,j=1,\ldots,\ell}$, wobei $\alpha_i = P(X_0 = i)$ und $p_{ij} = P(\varphi(i, Z_1) = j)$ für $i, j \in E$.

Bemerkung 6.6.2

1. Die Messbarkeitsbedingung, die die Abbildung φ im obigen Theorem erfüllen muss, ist rein technischer Natur. In allen Anwendungen und Beispielen, die in diesem Modul betrachtet werden, wird diese Bedingung stets erfüllt sein.

2. Da die Innovationen $Z_1, Z_2, \ldots$ identisch verteilt sind, gilt

$$p_{ij} = P(\varphi(i, Z_n) = j)$$

für jedes $n \geq 1$ und $i, j \in E$.

3. Umgekehrt kann man zeigen, dass es zu jeder endlichen Menge $E = \{1, \ldots, \ell\}$, jeder Wahrscheinlichkeitsfunktion $\boldsymbol{\alpha}$ auf E und jeder stochastischen $\ell \times \ell$ Matrix $\boldsymbol{P}$ eine Markov-Kette $X_0, X_1, \ldots$ mit Zustandsraum E, Anfangsverteilung $\boldsymbol{\alpha}$ und Übergangsmatrix $\boldsymbol{P}$ gibt, die eine rekursive Darstellung wie in obigem Theorem besitzt.

Beispiel 6.6.2 (*Zufällige Irrfahrten*) Klassische Beispiele für Markov-Ketten sind durch sogenannte zufällige Irrfahrten gegeben, die auch *Random Walk* genannt werden. Wir betrachten hier nur zufällige Irrfahrten mit dem ganzzahligen und beschränkten Wertebereich $E = \{-\ell, \ldots, 0, \ldots, \ell\}$ für ein $\ell \in \mathbb{N}$.

Es sei $Z_1, Z_2, \ldots : \Omega \to \mathbb{Z}$ eine Folge von unabhängigen und identisch verteilten Zufallsvariablen, die nur Werte in der Menge $\mathbb{Z} = \{\ldots, -1, 0, 1, \ldots\}$ der ganzen Zahlen annehmen. Weiterhin sei $X_0 : \Omega \to E$ eine beliebige Zufallsvariable, die von den „Innovationen" $Z_1, Z_2, \ldots$ unabhängig ist. Für $n \in \mathbb{N}$ sei die Zufallsvariable $X_n : \Omega \to E$ durch die folgende Rekursionsgleichung gegeben:

$$\begin{aligned} X_n &= \min\{\max\{X_{n-1} + Z_n, -\ell\}, \ell\} \\ &= \begin{cases} X_{n-1} + Z_n\,, & \text{wenn } -\ell \leq X_{n-1} + Z_n \leq \ell, \\ -\ell\,, & \text{wenn } X_{n-1} + Z_n < -\ell, \\ \ell\,, & \text{wenn } X_{n-1} + Z_n > \ell. \end{cases} \end{aligned}$$

Dann bilden die Zufallsvariablen $X_0, X_1, \ldots$ eine Markov-Kette mit dem Zustandsraum $E = \{-\ell, \ldots, 0, \ldots, \ell\}$, der Anfangsverteilung $\boldsymbol{\alpha} = (\alpha_{-\ell}, \ldots, \alpha_\ell)^\top$, wobei $\alpha_i = P(X_0 = i)$ für jedes $i \in E$, und mit der Übergangsmatrix

$\boldsymbol{P} = (p_{ij})$, die durch

$$p_{ij} = \begin{cases} P(Z_1 = j - i)\,, & \text{wenn } -\ell < j < \ell, \\ P(Z_1 \le j - i)\,, & \text{wenn } j = -\ell, \\ P(Z_1 \ge j - i)\,, & \text{wenn } j = \ell \end{cases}$$

für $i, j \in E$ gegeben ist.

Durch eine so gegebene zufällige Irrfahrt kann zum Beispiel die *Risikoreserve* von versicherungs- bzw. finanztechnischen Bilanzierungprozessen modelliert werden. Sei $X_0 : \Omega \to E$ die (zufällige) Anfangsreserve und $Z_n : \Omega \to \mathbb{Z}$ der „Zuwachs" in der n-ten Periode, der beispielsweise als Differenz $Z_n = a - Z'_n$ aus risikofreien Einnahmen $a > 0$ und zufallsbedingten Ausgaben/Verlusten Z'_n dargestellt werden kann. Außerdem nehmen wir an, dass zusätzliche Reserven durch Kredite bereitgestellt werden, wenn die Reserve unter $-\ell$ fällt und dass Reserven zur Tilgung anderer Verbindlichkeiten entnommen werden, falls diese den Wert ℓ überschreiten. Dann kann die Risikoreserve durch die Markov-Kette $X_0, X_1, \ldots$ wie oben beschrieben modelliert werden.

Bisher haben wir nur die Matrix $\boldsymbol{P} = (p_{ij})$ der ein-Schritt Übergangswahrscheinlichkeiten einer Markov-Kette $X_0, X_1, \ldots$ betrachtet, d.h., $p_{ij} = P(X_1 = j \,|\, X_0 = i)$ gibt die Wahrscheinlichkeit an, in einem (beliebigen) Zeitschritt vom Zustand i in den Zustand j zu gelangen, wenn $P(X_0 = i) > 0$. Analog dazu ist es manchmal sinnvoll, für ein beliebiges $n \in \mathbb{N}$ die Wahrscheinlichkeit zu betrachten, in genau n Schritten vom Zustand i in den Zustand j zu gelangen. Diese Wahrscheinlichkeit wird mit $p_{ij}^{(n)} = P(X_n = j \,|\, X_0 = i)$ für $P(X_0 = i) > 0$ bezeichnet. Die Matrix $\boldsymbol{P}^{(n)} = (p_{ij}^{(n)})_{i,j=1,\ldots,n}$ heißt *n-Schritt Übergangsmatrix* der Markov-Kette.

Die n-Schritt Übergangswahrscheinlichkeit $p_{ij}^{(n)}$ kann wie folgt bestimmt werden. Für beliebige $i_1, \ldots, i_{n-1} \in E$ kann das Produkt $p_{ii_1} p_{i_1 i_2} \cdots p_{i_{n-1} j}$ als die Wahrscheinlichkeit des „Pfades" $i \to i_1 \to i_2 \to \ldots \to i_{n-1} \to j$ aufgefasst werden, d.h., als die Wahrscheinlichkeit über die Zustände $i_1, \ldots, i_{n-1}$ in n Schritten von i nach j zu gelangen. Die Summe der Wahrscheinlichkeiten aller „Pfade" von i nach j in n Schritten ergibt dann $p_{ij}^{(n)}$:

$$p_{ij}^{(n)} = \sum_{i_1,\ldots,i_{n-1} \in E} p_{ii_1} p_{i_1 i_2} \cdots p_{i_{n-1} j}\,.$$

Die n-Schritt Übergangsmatrix $\boldsymbol{P}^{(n)}$ kann demzufolge als die n-te Potenz der (ein-Schritt) Übergangsmatrix $\boldsymbol{P}$ bestimmt werden, d.h., $\boldsymbol{P}^{(n)} = \boldsymbol{P}^n$ für jedes $n \in \mathbb{N}$. Insbesondere für große n und ℓ ist die Bestimmung von $\boldsymbol{P}^n$ jedoch rechnerisch aufwendig.

Beispiel 6.6.3 (*Produktivität eines Mitarbeiters*) Wir greifen das erste Beispiel aus Abschnitt 6.6.1 wieder auf. Darin wurde die Produktivität ei-

nes Mitarbeiters durch eine Markov-Kette $X_0, X_1, \ldots$ mit Zustandsraum $E = \{1, 2\}$, Anfangsverteilung $\boldsymbol{\alpha} = (0, 1)^\top$ und allgemeiner Übergangsmatrix

$$\boldsymbol{P} = \begin{pmatrix} 1-p & p \\ p' & 1-p' \end{pmatrix}$$

mit $p, p' \in (0, 1)$ modelliert. Mittels vollständiger Induktion kann man zeigen, dass die n-Schritt Übergangsmatrix $\boldsymbol{P}^{(n)} = \boldsymbol{P}^n$ gegeben ist durch

$$\boldsymbol{P}^n = \frac{1}{p+p'} \begin{pmatrix} p' & p \\ p' & p \end{pmatrix} + \frac{(1-p-p')^n}{p+p'} \begin{pmatrix} p & -p \\ -p' & p' \end{pmatrix}.$$

Außerdem kann man kann sich leicht überlegen, dass mit Hilfe der n-Schritt Übergangsmatrix $\boldsymbol{P}^{(n)}$ auch die (unbedingte) Wahrscheinlichkeitsfunktion $\boldsymbol{\alpha}_n = (\alpha_{n1}, \ldots, \alpha_{n\ell})^\top$ der diskreten Zufallsvariablen X_n bestimmt werden kann, wobei $\alpha_{ni} = P(X_n = i)$ für $i = 1, \ldots, \ell$. Es gilt stets, dass

$$\boldsymbol{\alpha}_n^\top = \boldsymbol{\alpha}^\top \boldsymbol{P}^{(n)} \qquad \text{für jedes } n \in \mathbb{N},$$

wobei $\boldsymbol{\alpha}$ die Anfangsverteilung der Markov-Kette bezeichnet.

6.6.3 Ergodizität von Markov-Ketten

Wenn die Anzahl ℓ der potentiell möglichen Zustände einer Markov-Kette $X_0, X_1, \ldots$ groß ist, dann kann es, wie schon in Abschnitt 6.6.2 erwähnt, schwierig sein, die n-Schritt Übergangsmatrix $\boldsymbol{P}^{(n)} = (p_{ij}^{(n)})$, sowie die (unbedingte) Wahrscheinlichkeitsfunktion $\boldsymbol{\alpha}_n$ von X_n zu berechnen (insbesondere für große n).

Man kann jedoch Bedingungen angeben, unter denen für jedes $j \in E$ die beiden Grenzwerte

$$\lim_{n\to\infty} p_{ij}^{(n)} = \lim_{n\to\infty} P(X_n = j \mid X_0 = i)$$

und

$$\lim_{n\to\infty} \alpha_{nj} = \lim_{n\to\infty} P(X_n = j)$$

existieren (und darüber hinaus gleich sind und nicht von i abhängen), sodass anstelle von $p_{ij}^{(n)}$ bzw. α_{nj} der Grenzwert $\pi_j = \lim_{n\to\infty} p_{ij}^{(n)} = \lim_{n\to\infty} \alpha_{nj}$ als *Näherungslösung* betrachtet werden kann, falls n groß ist.

Dies führt zu dem folgenden Begriff der *Ergodizität* von Markov-Ketten.

Definition 6.6.2 (Ergodizität) Die Markov-Kette $X_0, X_1, \ldots$ mit den n-Schritt Übergangsmatrizen $\boldsymbol{P}^{(n)} = (p_{ij}^{(n)})$ für $n \in \mathbb{N}$ heißt *ergodisch*, falls die Grenzwerte

$$\pi_j = \lim_{n\to\infty} p_{ij}^{(n)}$$

für jedes $j \in E$ existieren, positiv sind und nicht von $i \in E$ abhängen.

Bemerkung 6.6.3 1. Weil die Grenzverteilung $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_\ell)$ mit $\pi_j = \lim_{n\to\infty} p_{ij}^{(n)}$ einer ergodischen Markov-Kette $X_0, X_1, \ldots$ nicht von i abhängt, ergibt sich hieraus und aus der Endlichkeit des Zustandsraumes $E = \{1, \ldots, \ell\}$, dass

$$\lim_{n\to\infty} \boldsymbol{\alpha}_n^\top = \boldsymbol{\alpha}^\top \lim_{n\to\infty} \boldsymbol{P}^{(n)} = \boldsymbol{\pi}^\top .$$

Die positive Wahrscheinlichkeit π_j ist also der Grenzwert sowohl der bedingten Wahrscheinlichkeit $P(X_n = j \,|\, X_0 = i)$ für beliebiges $i \in E$ als auch der unbedingten Wahrscheinlichkeit $P(X_n = j)$ für jedes $j \in E$ und $n \to \infty$.

2. Die Grenzverteilung $\boldsymbol{\pi}$ kann demnach wie folgt interpretiert werden. Egal in welchem Zustand sich eine ergodische Markov-Kette zum Zeitpunkt $n = 0$ befindet, nach einer großen Anzahl n von Zeitschritten ist der zufällige Zustand X_n näherungsweise gemäß der Wahrscheinlichkeitsfunktion $\boldsymbol{\pi}$ verteilt. Dieses Resultat werden wir in Abschnitt 6.7 nutzen, um zufällige Strukturen mit Wahrscheinlichkeitsfunktion $\boldsymbol{\pi}$ mit Hilfe von ergodischen Markov-Ketten zu simulieren.

Beispiel 6.6.4 (*Produktivität eines Mitarbeiters)* Wir greifen noch einmal das Beispiel einer Markov-Kette $X_0, X_1, \ldots$ zur Modellierung der Produktivität eines Mitarbeiters auf, siehe Abschnitt 6.6.1 bzw. 6.6.2.

Die n-Schritt Übergangsmatrix $\boldsymbol{P}^{(n)} = \boldsymbol{P}^n$ ist in diesem Beispiel gegeben durch

$$\boldsymbol{P}^n = \frac{1}{p+p'} \begin{pmatrix} p' & p \\ p' & p \end{pmatrix} + \frac{(1-p-p')^n}{p+p'} \begin{pmatrix} p & -p \\ -p' & p' \end{pmatrix} ,$$

mit $p, p' \in (0,1)$. Da $|1-p-p'| < 1$, ergibt sich hieraus, dass

$$\lim_{n\to\infty} \boldsymbol{P}^n = \frac{1}{p+p'} \begin{pmatrix} p' & p \\ p' & p \end{pmatrix} ,$$

und damit

$$\boldsymbol{\pi} = \lim_{n\to\infty} \boldsymbol{\alpha}_n = \Big(\frac{p'}{p+p'} \,,\, \frac{p}{p+p'} \Big)^\top .$$

Die Grenzverteilung $\boldsymbol{\pi}$ lässt sich jedoch nur in Ausnahmefällen so direkt wie im vorhergehenden Beispiel bestimmen. Daher ist es oft auch nicht möglich, die Ergodizität einer Markov-Kette direkt mit Hilfe der Definition nachzuweisen. Im folgenden Theorem geben wir sowohl ein Kriterium an, um nachzuweisen, dass eine Markov-Kette ergodisch ist, als auch eine Möglichkeit, die eindeutig bestimmte Grenzverteilung $\boldsymbol{\pi}$ durch das Lösen eines linearen Gleichungssystems zu ermitteln.

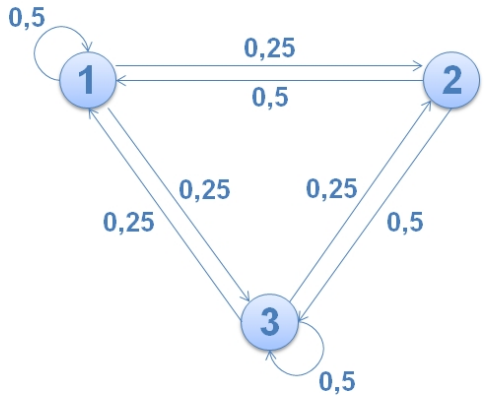


Abbildung 6.2: Beispielhafte Darstellung der Übergangsmatrix einer Markov-Kette als geometrischer Graph mit gewichteten und gerichteten Kanten.

Satz 6.6.2

1. Die Markov-Kette $X_0, X_1, \ldots$ mit dem Zustandsraum $E = \{1, \ldots, \ell\}$ und der Übergangsmatrix $\boldsymbol{P}$ ist genau dann ergodisch, wenn es eine natürliche Zahl $n_0 \in \mathbb{N}$ gibt, sodass $\boldsymbol{P}^{n_0}$ eine positive Matrix ist, d.h., sodass alle Einträge von $\boldsymbol{P}^{n_0}$ positiv sind. Die Übergangsmatrix $\boldsymbol{P}$ wird dann *quasi-positiv* genannt.

2. Sei $X_0, X_1, \ldots$ eine ergodische Markov-Kette mit dem Zustandsraum $E = \{1, \ldots, \ell\}$ und der Übergangsmatrix $\boldsymbol{P} = (p_{ij})$. Dann ist die Grenzverteilung $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_\ell)^\top$ gleichzeitig die eindeutig bestimmte (positive) Lösung des linearen Gleichungssystems

$$\pi_j = \sum_{i \in E} \pi_i p_{ij}\,, \qquad j \in E,$$

(bzw. $\boldsymbol{\pi}^\top = \boldsymbol{\pi}^\top \boldsymbol{P}$ in Matrixschreibweise), die der Normierungsbedingung $\sum_{j \in E} \pi_j = 1$ genügt.

Bemerkung 6.6.4 Ohne die zusätzliche Normierungsbedingung $\sum_{j \in E} \pi_j = 1$ ist das lineare Gleichungssystem im vorhergehenden Theorem nicht eindeutig lösbar, da die zugehörige Koeffizientenmatrix keinen vollen Rang besitzt.

Beispiel 6.6.5 (*Übergangsmatrix als Graph*) Wenn die Anzahl ℓ von potenziellen Zuständen einer Markov-Kette $X_0, X_1, \ldots$ mit Zustandsraum E nicht zu groß ist, dann kann man die zugehörige Übergangsmatrix $\boldsymbol{P} = (p_{ij})$ als geometrischen Graphen mit gewichteten und gerichteten Kanten interpretieren. Jeder Zustand $i \in E$ wird durch einen Knoten dargestellt und für $i, j \in E$ existiert eine gerichtete Kante von Knoten i zu Knoten j, ge-

nau dann wenn $p_{ij} > 0$, wobei diese Kante mit der Wahrscheinlichkeit p_{ij} gewichtet wird.

Wir betrachten eine Markov-Kette $X_0, X_1, \ldots$ mit Zustandsraum $E = \{1, 2, 3\}$ und Übergangsmatrix

$$\boldsymbol{P} = \begin{pmatrix} 0,5 & 0,25 & 0,25 \\ 0,5 & 0 & 0,5 \\ 0,25 & 0,25 & 0,5 \end{pmatrix}.$$

Die Übergangsmatrix lässt sich auch durch den Graphen in Abbildung 6.2 darstellen. Außerdem gilt

$$\boldsymbol{P}^2 = \begin{pmatrix} 0,4375 & 0,1875 & 0,375 \\ 0,375 & 0,25 & 0,375 \\ 0,375 & 0,1875 & 0,4375 \end{pmatrix},$$

d.h., $\boldsymbol{P}$ ist quasi-positiv und die Markov-Kette $X_0, X_1, \ldots$ ist somit ergodisch.

Wir bestimmen die eindeutige Grenzverteilung $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)^\top$ wie im vorhergehenden Theorem beschrieben. Gesucht ist die Lösung des Gleichungssystems

$$\begin{array}{lll} \text{I.} & \pi_1 = 0,5\pi_1 + 0,5\pi_2 + 0,25\pi_3 & \\ \text{II.} & \pi_2 = 0,25\pi_1 + 0,25\pi_3 & \Leftrightarrow \\ \text{III.} & 1 = \pi_1 + \pi_2 + \pi_3 & \end{array} \qquad \begin{array}{ll} \text{I.} & 0 = -0,5\pi_1 + 0,5\pi_2 + 0,25\pi_3 \\ \text{II.} & 0 = 0,25\pi_1 - \pi_2 + 0,25\pi_3 \\ \text{III.} & 1 = \pi_1 + \pi_2 + \pi_3 \end{array}$$

Aus 2I. + II. folgt $0 = -0,75\pi_1 + 0,75\pi_3$ und daraus wiederum $\pi_1 = \pi_3$. Aus II. + III. hingegen folgt $1 = 1,25\pi_1 + 1,25\pi_3 = 2,5\pi_1$. Daher gilt $\pi_1 = \pi_3 = 0,4$ und somit $\pi_2 = 0,2$. Die eindeutig bestimmte Grenzverteilung der ergodischen Markov-Kette $X_0, X_1, \ldots$ ist also gegeben durch $\boldsymbol{\pi} = (0,4 \;\; 0,2 \;\; 0,4)^\top$.

In vielen Anwendungen werden jedoch ergodische Markov-Ketten mit einem viel größeren Zustandsraum als im vorhergehenden Beispiel betrachtet. Wenn dann auch noch die Übergangsmatrix $\boldsymbol{P}$ dünn besetzt ist, d.h., wenn viele Einträge gleich null sind, ist es schwer, zu überprüfen, ob die Übergangsmatrix $\boldsymbol{P}$ quasi-positiv ist. Wir führen deshalb noch eine andere (probabilistische) Charakterisierung der Ergodizität ein, die auf der *Irreduzibilität* und *Aperiodizität* von Markov-Ketten beruht. Diese Begriffe werden im Folgenden erläutert.

Definition 6.6.3 (Erreichbarkeit/Irreduzibilität) Betrachte eine Markov-Kette $X_0, X_1, \ldots$ mit Zustandsraum E, Übergangsmatrix $\boldsymbol{P} = (p_{ij})$ und n-Schritt Übergangsmatrizen $\boldsymbol{P}^{(n)} = (p_{ij}^{(n)})$ für $n \in \mathbb{N}$.

1. Für beliebige Zustände $i, j \in E$ sagen wir, dass der Zustand j vom Zustand i *erreichbar* ist, wenn $p_{ij}^{(n)} > 0$ für ein $n \in \mathbb{N}$ (Schreibweise: $i \to j$).

2. Man sagt, dass die Zustände $i, j \in E$ *kommunizieren*, wenn $i \to j$ und $j \to i$ (Schreibweise $i \leftrightarrow j$).

3. Die Markov-Kette $X_0, X_1, \ldots$ bzw. ihre Übergangsmatrix $\boldsymbol{P}$ heißt *irreduzibel*, wenn jeder Zustand von jedem Zustand aus erreichbar ist, d.h., wenn $i \leftrightarrow j$ für alle $i, j \in E$.

Definition 6.6.4 (Periode/Aperiodizität) Es sei $X_0, X_1, \ldots$ eine Markov-Kette mit Zustandsraum E, Übergangsmatrix $\boldsymbol{P} = (p_{ij})$ und n-Schritt Übergangsmatrizen $\boldsymbol{P}^{(n)} = (p_{ij}^{(n)})$ für $n \in \mathbb{N}$.

1. Die *Periode* d_i des Zustandes $i \in E$ ist definiert als $d_i = \text{ggt}\{n \in \mathbb{N} : p_{ii}^{(n)} > 0\}$, wobei „ggt" den größten gemeinsamen Teiler bezeichnet. Dabei setzen wir $d_i = \infty$, falls $p_{ii}^{(n)} = 0$ für jedes $n \in \mathbb{N}$.

2. Der Zustand $i \in E$ heißt *aperiodisch*, wenn $d_i = 1$.

3. Die Markov-Kette $X_0, X_1, \ldots$ bzw. ihre Übergangsmatrix $\boldsymbol{P} = (p_{ij})$ heißt *aperiodisch*, wenn sämtliche Zustände von $X_0, X_1, \ldots$ aperiodisch sind.

Bemerkung 6.6.5

1. Wenn $p_{ii} > 0$, dann ist der Zustand i stets aperiodisch.

2. Mann kann zeigen, dass $d_i = d_j$, wenn die Zustände $i, j \in E$ kommunizieren. Insbesondere haben alle Zustände einer irreduziblen Markov-Kette die gleiche Periode. Möchte man also nachweisen, dass eine irreduzible Markov-Kette aperiodisch ist, so genügt es zu zeigen, dass ein (beliebig gewählter) Zustand $i \in E$ aperiodisch ist (z.B., wenn $p_{ii} > 0$ für ein $i \in E$).

Die Ergodizität einer Markov-Kette lässt sich nun wie folgt charakterisieren.

Satz 6.6.3 Eine Markov-Kette $X_0, X_1, \ldots$ mit endlichem Zustandsraum $E = \{1, \ldots, \ell\}$ ist genau dann ergodisch, wenn sie irreduzibel und aperiodisch ist.

Daher lässt sich in vielen Fällen die Ergodizität einer Markov-Kette am einfachsten nachweisen, indem man zeigt, dass die Markov-Kette irreduzibel und aperiodisch ist. Dies wird in den folgenden Beispielen veranschaulicht.

Beispiel 6.6.6 1. *Übergangsmatrix als Graph*

Wir kehren zunächst nochmals zu der im letzten Beispiel betrachteten Markov-Kette $X_0, X_1, \ldots$ zurück, deren Übergangsmatrix in Abbildung 6.2 dargestellt ist. Wir haben bereits gezeigt, dass $X_0, X_1, \ldots$ ergodisch und daher auch irreduzibel und aperiodisch ist. Alternativ lassen sich die Eigenschaften der Irrreduzibilität und der Aperiodizität von $X_0, X_1, \ldots$ auch direkt mit Hilfe der entsprechenden Definitionen zeigen. Weil $p_{ij}^{(2)} > 0$ und damit $i \to j$ für alle $i, j \in E$, ist die Markov-Kette irreduzibel. Weil $p_{11} > 0$, gilt $d_1 = 1$ und somit auch $d_2 = d_3 = 1$. Dementsprechend ist $X_0, X_1, \ldots$ ebenfalls aperiodisch.

2. *Zufällige Irrfahrten*

   Wir greifen noch einmal das in Abschnitt 6.6.2 betrachtete Beispiel über zufällige Irrfahrten auf. Die Markov-Kette $X_0, X_1, \ldots$ mit Zustandsraum $E = \{-\ell, \ldots, 0, \ldots, \ell\}$ für ein $\ell \in \mathbb{N}$ ist rekursiv definiert durch $X_n = \min\{\max\{X_{n-1} + Z_n, -\ell\}, \ell\}$ für $n \in \mathbb{N}$, wobei X_0 eine beliebige Zufallsvariable mit Werten in E und $Z_1, Z_2, \ldots$ eine Folge von unabhängigen und identisch verteilten Zufallsvariablen („Innovationen") mit Werten in $\mathbb{Z}$ ist, die außerdem unabhängig von X_0 sind.

   Man kann sich leicht überlegen, dass es im Allgemeinen von der Beschaffenheit der Verteilung der Innovationen abhängt, ob die Markov-Kette $X_0, X_1, \ldots$ ergodisch ist oder nicht. Wenn z.B. $P(Z_1 = k) = 0$ für jedes $k \in \{0, 1, \ldots\}$, dann ist der Zustand ℓ nicht vom Zustand 0 erreichbar, weil der Zustand nie größer werden kann. In diesem Fall ist die Markov-Kette $X_0, X_1, \ldots$ nicht irreduzibel und daher auch nicht ergodisch.

   Gilt hingegen $P(Z_1 = 1) > 0$ und $P(Z_1 = -1) > 0$, dann ist $X_0, X_1, \ldots$ irreduzibel, da jeder Zustand von jedem anderen Zustand in endlich vielen Schritten erreichbar ist und somit alle Zustände $i, j \in E$ kommunizieren. In diesem Fall gilt dann auch, dass $p_{\ell\ell} > 0$, da die Markov-Kette im Zustand ℓ verbleibt, wenn eine positive Innovation folgt (was mit positiver Wahrscheinlichkeit geschieht). Somit ist $d_\ell = 1$ und, da $X_0, X_1, \ldots$ irreduzibel ist, gilt, dass $d_i = 1$ für jeden Zustand $i \in E$. Die Markov-Kette $X_0, X_1, \ldots$ ist demnach auch aperiodisch und folglich ergodisch.

3. *Diffusionsmodell nach Ehrenfest*

   Wir betrachten das in Abschnitt 6.6.1 eingeführte Diffusionsmodell. Dieses ist durch eine Markov-Kette $X_0, X_1, \ldots$ mit Zustandsraum $E = \{0, \ldots, \ell\}$, beliebiger Anfangsverteilung $\boldsymbol{\alpha}$ und Übergangsmatrix $\boldsymbol{P} =$

(p_{ij}) gegeben, wobei

$$p_{ij} = \begin{cases} \dfrac{\ell - i}{\ell}\,, & \text{falls } i < \ell \text{ und } j = i + 1, \\ \dfrac{i}{\ell}\,, & \text{falls } i > 0 \text{ und } j = i - 1, \\ 0\,, & \text{sonst.} \end{cases}$$

Man kann sich leicht überlegen, dass $X_0, X_1, \ldots$ irreduzibel ist. Seien nämlich $i, j \in E$ beliebige Zustände und sei ohne Beschränkung der Allgemeinheit $i < j$. Dann gilt, dass $i \to i+1 \to \ldots \to j-1 \to j$, da stets eine positive Wahrscheinlichkeit besteht, dass sich der Zustand erhöht und somit $p_{ij}^{(j-i)} > 0$. Analog gilt $j \to j-1 \to \ldots \to i+1 \to i$ und demnach $p_{ji}^{(j-i)} > 0$. Die Zustände i und j kommunizieren also und da sie beliebig gewählt wurden, ist die Markov-Kette $X_0, X_1, \ldots$ irreduzibel.

Man kann sich jedoch leicht überlegen, dass $X_0, X_1, \ldots$ nicht aperiodisch und daher auch nicht ergodisch ist. Die Markov-Kette benötigt stets eine gerade Anzahl an (Zeit-) Schritten, um von einem Zustand $i \in E$ zu diesem Zustand zurück zu kehren. Es gilt deshalb, dass $d_i = \operatorname{ggt}\{n \in \mathbb{N} : p_{ii}^{(n)} > 0\} = 2$ für jedes $i \in E$.

6.6.4 Stationäre Anfangsverteilung und Reversibilität

Wir betrachten eine beliebige Markov-Kette $X_0, X_1, \ldots$ mit Zustandsraum $E = \{1, \ldots, \ell\}$ und Übergangsmatrix $\boldsymbol{P}$. Im Allgemeinen ist es möglich, dass das lineare Gleichungssystem $\boldsymbol{\alpha}^\top = \boldsymbol{\alpha}^\top \boldsymbol{P}$ eine oder mehrere Lösungen $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_\ell)^\top$ mit $\sum_{i \in E} \alpha_i = 1$ besitzt, wobei der letzte Fall nur dann eintreten kann, wenn die Markov-Kette $X_0, X_1, \ldots$ nicht ergodisch ist. Wenn die Anfangsverteilung von $X_0, X_1, \ldots$ durch eine solche Wahrscheinlichkeitsfunktion $\boldsymbol{\alpha}$ gegeben ist, dann gilt für jedes $n \in \mathbb{N}$, dass

$$\boldsymbol{\alpha}_n^\top = \boldsymbol{\alpha}^\top \boldsymbol{P}^n = \boldsymbol{\alpha}^\top \boldsymbol{P} \boldsymbol{P}^{(n-1)} = \boldsymbol{\alpha}^\top \boldsymbol{P}^{(n-1)} = \ldots = \boldsymbol{\alpha}^\top \boldsymbol{P} = \boldsymbol{\alpha}^\top\,,$$

wobei $\boldsymbol{\alpha}_n$ die in Abschnitt 6.6.2 eingeführte Wahrscheinlichkeitsfunktion von X_n ist. Die Zufallsvariablen $X_0, X_1, \ldots$ sind also in diesem Fall identisch verteilt und die Wahrscheinlichkeitsfunktion $\boldsymbol{\alpha}$ wird dann *stationäre Anfangsverteilung* der Markov-Kette $X_0, X_1, \ldots$ genannt.

Bemerkung 6.6.6 Im Allgemeinen darf die stationäre Anfangsverteilung einer Markov-Kette nicht mit der in Abschnitt 6.6.3 eingeführten Grenzverteilung $\boldsymbol{\pi} = \lim_{n \to \infty} \boldsymbol{\alpha}_n$ verwechselt werden. Im Folgenden wird der Unterschied zwischen diesen beiden Begriffen nochmals genauer erläutert.

1. Wenn die Markov-Kette $X_0, X_1, \ldots$ ergodisch ist, so stellt die Grenzverteilung $\boldsymbol{\pi}$ die einzige stationäre Anfangsverteilung der Markov-Kette

dar. Hat die Markov-Kette $X_0, X_1, \ldots$ eine beliebige Anfangsverteilung, dann ist, gemäß der Definition der Ergodizität, die Zufallsvariable X_n für große n näherungsweise gemäß $\boldsymbol{\pi}$ verteilt. Ist die Wahrscheinlichkeitsfunktion $\boldsymbol{\pi}$ hingegen auch die Anfangsverteilung der Markov-Kette, so ist X_n sogar für alle $n \geq 0$ gemäß $\boldsymbol{\pi}$ verteilt.

2. Wenn die Markov-Kette $X_0, X_1, \ldots$ irreduzibel aber nicht aperiodisch (und damit auch nicht ergodisch) ist, dann existiert genau eine stationäre Anfangsverteilung $\boldsymbol{\alpha}$. Hat $X_0, X_1, \ldots$ also die stationäre Anfangsverteilung $\boldsymbol{\alpha}$, so ist X_n für alle $n \geq 0$ gemäß $\boldsymbol{\alpha}$ verteilt.

3. Wenn die Markov-Kette $X_0, X_1, \ldots$ nicht irreduzibel ist, dann gibt es im Allgemeinen sogar mehrere stationäre Anfangsverteilungen, da das am Anfang dieses Abschnitts aufgestellte Gleichungssystem nicht eindeutig lösbar ist. Die Interpretation der stationären Anfangsverteilungen ergibt sich analog zu Punkt 2.

Beispiel 6.6.7 (*Diffusionsmodell nach Ehrenfest*) Wir greifen wieder das in den Abschnitten 6.6.1 und 6.6.3 eingeführte Diffusionsmodell auf. Es wurde bereits gezeigt, dass die betrachtete Markov-Kette $X_0, X_1, \ldots$ irreduzibel, aber nicht aperiodisch und somit auch nicht ergodisch ist. Die Grenzverteilung $\boldsymbol{\pi}$ existiert also nicht, allerdings besitzt die Markov-Kette eine eindeutig bestimmte stationäre Anfangsverteilung $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_\ell)^\top$, die durch $\alpha_j = \frac{1}{2^\ell}\binom{\ell}{j}$ für jedes $j \in \{0, \ldots, \ell\}$ gegeben ist, d.h., $\boldsymbol{\alpha}$ ist die Wahrscheinlichkeitsfunktion einer Binomialverteilung mit Parametern ℓ und $\frac{1}{2}$. Dies kann wie folgt gezeigt werden.

Da $\boldsymbol{\alpha}$ eine Wahrscheinlichkeitsfunktion ist, gilt offensichtlich $\sum_{j\in E} \alpha_j = 1$. Sei zunächst $j = 0$. Dann gilt

$$\sum_{i\in E} \alpha_i p_{i0} \;=\; \frac{1}{\ell}\alpha_1 \;=\; \frac{1}{\ell}\frac{1}{2^\ell}\ell \;=\; \frac{1}{2^\ell} \;=\; \alpha_0.$$

Ebenso gilt für $j = \ell$

$$\sum_{i\in E} \alpha_i p_{i\ell} \;=\; \frac{1}{\ell}\alpha_{\ell-1} \;=\; \frac{1}{\ell}\frac{1}{2^\ell}\binom{\ell}{\ell-1} \;=\; \frac{1}{\ell}\frac{1}{2^\ell}\ell \;=\; \frac{1}{2^\ell} \;=\; \alpha_\ell.$$

Sei schließlich $j \in \{1, \ldots, \ell-1\}$. Dann ergibt sich

$$\begin{aligned}
&\sum_{i\in E} \alpha_i p_{ij} \;=\; \alpha_{j-1}\frac{\ell-(j-1)}{\ell} + \alpha_{j+1}\frac{j+1}{\ell} \\
&= \frac{1}{2^\ell}\binom{\ell}{j-1}\frac{\ell-(j-1)}{\ell} + \frac{1}{2^\ell}\binom{\ell}{j+1}\frac{j+1}{\ell} \\
&= \frac{1}{2^\ell}\left(\frac{\ell!}{(j-1)!\,(\ell-(j-1))!}\,\frac{\ell-(j-1)}{\ell} + \frac{\ell!}{(j+1)!\,(\ell-(j+1))!}\,\frac{j+1}{\ell}\right)
\end{aligned}$$

$$= \frac{1}{2^\ell}\left(\frac{(\ell-1)!}{(j-1)!\,(\ell-j)!} + \frac{(\ell-1)!}{j!\,(\ell-1-j)!}\right) = \frac{1}{2^\ell}\,\frac{(\ell-1)!\,j + (\ell-1)!\,(\ell-j)}{j!\,(\ell-j)!}$$
$$= \frac{1}{2^\ell}\,\frac{\ell!}{j!\,(\ell-j)!} = \frac{1}{2^\ell}\binom{\ell}{j} = \alpha_j\,.$$

Somit ist $\boldsymbol{\alpha}$ die eindeutig bestimmte stationäre Anfangsverteilung der Markov-Kette $X_0, X_1, \ldots$. Wenn also $\boldsymbol{\alpha}$ die Anfangsverteilung der Markov-Kette ist, dann ist $X_n \sim \mathrm{Bin}(\ell, \frac{1}{2})$ für jedes $n \in \mathbb{N}$. Für eine beliebige Anfangsverteilung ist das im Allgemeinen nicht der Fall.

Wir wollen uns nun noch mit dem Begriff der Reversibilität einer Markov-Kette $X_0, X_1, \ldots$ beschäftigen. Dieser bietet eine weitere Möglichkeit, um zu überprüfen, ob eine Wahrscheinlichkeitsfunktion $\boldsymbol{\alpha}$ eine stationäre Anfangsverteilung (und somit im ergodischen Fall die eindeutig bestimmte Grenzverteilung) von $X_0, X_1, \ldots$ ist.

Definition 6.6.5 (Reversibilität) Sei $X_0, X_1, \ldots$ eine Markov-Kette mit Zustandsraum $E = \{1, \ldots, \ell\}$ und Übergangsmatrix $\boldsymbol{P} = (p_{ij})$ und sei $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_\ell)^\top$ eine beliebige Wahrscheinlichkeitsfunktion auf E, sodass $\alpha_i > 0$ für jedes $i \in E$. Die Markov-Kette $X_0, X_1, \ldots$ bzw. das Paar $(\boldsymbol{P}, \boldsymbol{\alpha})$ heißt *reversibel*, wenn $\alpha_i\, p_{ij} = \alpha_j\, p_{ji}$ für beliebige $i, j \in E$. Diese Bedingung wird auch *Detailed-Balance-Bedingung* genannt.

Bemerkung 6.6.7 Man kann leicht zeigen, dass die Wahrscheinlichkeitsfunktion $\boldsymbol{\alpha}$ eine stationäre Anfangsverteilung der Markov-Kette $X_0, X_1, \ldots$ darstellt, wenn $(\boldsymbol{P}, \boldsymbol{\alpha})$ reversibel ist. Insbesondere ist $\boldsymbol{\alpha}$ identisch mit der eindeutig bestimmten Grenzverteilung $\boldsymbol{\pi} = \lim_{n\to\infty} \boldsymbol{\alpha}_n$, wenn $X_0, X_1, \ldots$ ergodisch ist. Aus der Detailed-Balance-Bedingung in der Definition der Reversibilität folgt also stets die „globale" Balance-Bedingung $\boldsymbol{\alpha}^\top = \boldsymbol{\alpha}^\top \boldsymbol{P}$.

Beispiel 6.6.8 (*Diffusionsmodell nach Ehrenfest*) Wir kehren ein weiteres Mal zu dem Diffusionsmodell zurück, welches bereits in den Abschnitten 6.6.1 und 6.6.3, sowie in diesem Abschnitt betrachtet wurde. Wir hatten bereits gezeigt, dass die Markov-Kette $X_0, X_1, \ldots$ mit Zustandsraum $E = \{0, \ldots, \ell\}$ und Übergangsmatrix $\boldsymbol{P} = (p_{ij})$, wobei

$$p_{ij} = \begin{cases} \dfrac{\ell - i}{\ell}\,, & \text{falls } i < \ell \text{ und } j = i+1, \\ \dfrac{i}{\ell}\,, & \text{falls } i > 0 \text{ und } j = i-1, \\ 0\,, & \text{sonst} \end{cases}$$

irreduzibel, aber nicht aperiodisch ist und dass die eindeutig bestimmte stationäre Anfangsverteilung $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_\ell)^\top$ durch $\alpha_i = \frac{1}{2^\ell}\binom{\ell}{i}$ für $i \in E$ gegeben ist.

Wenn $i < \ell$ und $j = i+1$, dann ist $j > 0$ und $i = j - 1$ und somit gilt,

dass

$$\begin{aligned}\alpha_i p_{ij} &= \frac{1}{2^\ell}\binom{\ell}{i}\frac{\ell-i}{\ell} = \frac{1}{2^\ell}\binom{\ell}{j-1}\frac{\ell-j+1}{\ell} \\ &= \frac{1}{2^\ell}\frac{\ell!}{(j-1)!\,(\ell-j+1)!}\frac{\ell-j+1}{\ell}\frac{j}{j} = \frac{1}{2^\ell}\binom{\ell}{j}\frac{j}{\ell} = \alpha_j p_{ji}\,.\end{aligned}$$

Eine analoge Rechnung ergibt sich, wenn $i > 0$ und $j = i - 1$. Da schließlich $p_{ij} = 0$ genau dann, wenn $p_{ji} = 0$, folgt, dass das Paar $(\boldsymbol{P}, \boldsymbol{\alpha})$ reversibel ist.

6.7 Markov-Chain-Monte-Carlo-Simulation

Im letzten Abschnitt dieses Kapitels wollen wir eine Klasse von dynamischen Simulationsalgorithmen kennen lernen, die auf der Konstruktion von Markov-Ketten basieren und auch in der deutschsprachigen Literatur *Markov Chain Monte Carlo (MCMC)* genannt werden.

Die allgemeine Vorgehensweise lässt sich wie folgt skizzieren. Gegeben sei ein endlicher Zustandsraum E, wobei die Elemente von E eine fast beliebig komplizierte Struktur haben können. Zum Beispiel kann E durch die Menge aller Graustufenbilder $\boldsymbol{x} = \{x(v), v \in V\}$ gegeben sein, wobei V eine endliche Menge an Pixeln ist und jedem Pixel $v \in V$ ein Graustufenwert $x(v) \in \{0, \ldots, 255\}$ zugewiesen wird. Des Weiteren sei $\boldsymbol{\pi} : E \to (0, 1)$ eine beliebige Wahrscheinlichkeitsfunktion auf E mit $\pi_{\boldsymbol{x}} > 0$ für jedes $\boldsymbol{x} \in E$ und $\sum_{\boldsymbol{x}\in E} \pi_{\boldsymbol{x}} = 1$. Wenn die Anzahl $|E|$ von Elementen in E groß ist, dann sind herkömmliche Methoden, wie z.B. die in Abschnitt 6.2 diskutierte Inversionsmethode, zur Generierung von Pseudozufallszahlen $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ aus E gemäß der Wahrscheinlichkeitsfunktion $\boldsymbol{\pi}$ im Allgemeinen ineffizient und daher nicht geeignet.

Ein alternativer Ansatz besteht darin, eine ergodische Markov-Kette $\boldsymbol{X}_0$, $\boldsymbol{X}_1, \ldots$ mit Zustandsraum E und geeigneter Übergangsmatrix $\boldsymbol{P}$ zu konstruieren, sodass $\boldsymbol{\pi}$ die Grenzverteilung dieser Markov-Kette ist. Dann ist für hinreichend große n die Zufallsvariable $\boldsymbol{X}_n$ näherungsweise gemäß $\boldsymbol{\pi}$ verteilt. Pseudozufallszahlen $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ in E mit Wahrscheinlichkeitsfunktion $\boldsymbol{\pi}$ können also effizient durch wiederholte Simulation der Markov-Kette generiert werden.

6.7.1 Gibbs-Sampler

Eine wichtige Klasse von MCMC-Algorithmen bilden sogenannte *Gibbs-Sampler* (nach Josiah Willard Gibbs, 1839-1903). Es sei V eine endliche Indexmenge und $\boldsymbol{X} = (X(v), v \in V)$ ein diskreter Zufallsvektor mit endlichem Wertebereich $E \subset \mathbb{R}^{|V|}$. Wir nehmen an, dass es für jedes Paar $\boldsymbol{x}, \boldsymbol{x}' \in E$

eine endliche Folge von Zuständen $\boldsymbol{y}_0, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \in E$ gibt, sodass

$$\boldsymbol{y}_0 = \boldsymbol{x}\,, \quad \boldsymbol{y}_n = \boldsymbol{x}'$$

und

$$\#\{v \in V : y_i(v) \neq y_{i+1}(v)\} = 1$$

für jedes $i = 0, \ldots, n-1$, d.h., jeder Zustand $\boldsymbol{y}_i$ der Folge $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ unterscheidet sich in genau einer Komponente von seinem Vorgänger.

Des Weiteren sei $\boldsymbol{\pi} = (\pi_{\boldsymbol{x}}, \boldsymbol{x} \in E)$ die Wahrscheinlichkeitsfunktion des Zufallsvektors $\boldsymbol{X}$, sodass $\pi_{\boldsymbol{x}} > 0$ für jeden Zustand $\boldsymbol{x} \in E$. Für jede Komponente $v \in V$ und jeden Zustand $\boldsymbol{x} \in E$ sei außerdem

$$\pi_{x(v)|\,\boldsymbol{x}(-v)} = P\big(X(v) = x(v) \mid \boldsymbol{X}(-v) = \boldsymbol{x}(-v)\big) = \frac{\pi_{(x(v),\boldsymbol{x}(-v))}}{\sum\limits_{\boldsymbol{y}\in E:\boldsymbol{y}(-v)=\boldsymbol{x}(-v)} \pi_{\boldsymbol{y}}}$$

die bedingte Wahrscheinlichkeit, dass die Komponente $X(v)$ von $\boldsymbol{X}$ den Wert $x(v)$ annimmt, unter der Bedingung, dass der Vektor $\boldsymbol{X}(-v) = (X(w), w \in V \setminus \{v\})$ aller anderen Komponenten $w \neq v$ gleich $\boldsymbol{x}(-v) = (x(w), w \in V \setminus \{v\})$ ist. Dabei setzen wir voraus, dass $(x(v), \boldsymbol{x}(-v)) \in E$. Schließlich sei $\boldsymbol{q} = (q_v, v \in V)$ eine Wahrscheinlichkeitsfunktion auf V, sodass $q_v > 0$ für jedes $v \in V$ und $\sum_{v\in V} q_v = 1$. Dann kann eine geeignete ergodische Markov-Kette wie folgt zur Simulation von $\boldsymbol{\pi}$ konstruiert werden.

Satz 6.7.1 Es sei $\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots$ eine Markov-Kette mit Zustandsraum E, beliebiger Anfangsverteilung $\boldsymbol{\alpha}$ und Übergangsmatrix $\boldsymbol{P} = (p_{\boldsymbol{x}\boldsymbol{x}'})$, die wie folgt gegeben ist:

$$p_{\boldsymbol{x}\boldsymbol{x}'} = \sum_{v\in V} q_v \pi_{x'(v)|\,\boldsymbol{x}(-v)}\, I_{\{\boldsymbol{x}(-v)=\boldsymbol{x}'(-v)\}} \qquad \text{für alle } \boldsymbol{x}, \boldsymbol{x}' \in E\,.$$

Dann ist $\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots$ irreduzibel und aperiodisch und das Paar $(\boldsymbol{P}, \boldsymbol{\pi})$ ist reversibel. Hieraus folgt insbesondere, dass $\boldsymbol{\pi}$ die eindeutig bestimmte Grenzverteilung der Markov-Kette $\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots$ ist.

Dieses Theorem besagt also, dass der Zufallsvektor $\boldsymbol{X}_n$ für große n näherungsweise gemäß der Wahrscheinlichkeitsfunktion $\boldsymbol{\pi}$ verteilt ist. Dementsprechend können Pseudozufallsvektoren mit Wahrscheinlichkeitsfunktion $\boldsymbol{\pi}$ wie folgt erzeugt werden.

Simulationsalgorithmus

1. Wähle einen beliebigen (Anfangs-) Zustand $\boldsymbol{x}_0 \in E$ und setze $n = 0$.
2. Generiere eine Pseudozufallszahl $v \in V$ gemäß der Wahrscheinlichkeitsfunktion $\boldsymbol{q}$, d.h., wähle eine zufällige Komponente in V.

3. Generiere eine Pseudozufallszahl x gemäß der bedingten Wahrscheinlichkeitsfunktion

$$\boldsymbol{\pi}_{\cdot\,|\boldsymbol{x}_n(-v)} = \big(\pi_{x(v)|\,\boldsymbol{x}_n(-v)},\ \text{für die } x(v),\ \text{für die } (x(v), \boldsymbol{x}_n(-v)) \in E\big)\,.$$

Setze $x_{n+1}(v) = x$.

4. Die Werte für alle anderen Komponenten $w \neq v$ werden nicht geändert, d.h., $x_{n+1}(w) = x_n(w)$ für jedes $w \in V \setminus \{v\}$. Setze $n = n + 1$.

5. Wenn n groß genug ist, dann akzeptiere $\boldsymbol{x}_n$ als Pseudozufallsvektor der Wahrscheinlichkeitsfunktion $\boldsymbol{\pi}$. Ansonsten geh zurück zu Punkt 2.

Beispiel 6.7.1 (*Hard-Core-Modell für soziale Netzwerke*) Die Mitarbeiterstruktur eines Unternehmens werde durch ein soziales Netzwerk modelliert. Jeder Mitarbeiter wird durch einen Knoten des Netzwerks dargestellt und die (endliche) Menge aller Knoten mit V bezeichnet. Zwei unterschiedliche Knoten sind durch eine Kante verbunden, wenn die zugehörigen Mitarbeiter an einem gemeinsamen Projekt arbeiten, wobei jeder Mitarbeiter in mehreren Projekten tätig sein kann.

Das Unternehmen möchte einige Mitarbeitern zu Projektleitern befördern. Wenn ein Mitarbeiter zum Projektleiter ernannt wird, dann wird der zugehörige Knoten im Netzwerk durch eine 1 markiert, ansonsten durch eine 0. Allerdings soll es pro Projekt maximal einen Projektleiter geben, d.h., dass zwei durch eine Kante verbundene Knoten nicht gleichzeitig mit 1 markiert sein können. Wir betrachten demzufolge den Zustandsraum $E \subset \{0,1\}^{|V|}$, der alle zulässigen Konfigurationen des Netzwerks enthält, d.h., alle Konfigurationen, bei denen es höchstens einen Projektleiter pro Projekt gibt (also keine zwei mit 1 markierten verbundenen Knoten). Im Allgemeinen wird ein solches Modell als Hard-Core-Modell bezeichnet.

Da alle Mitarbeiter vergleichbare Qualifikationen haben, möchte die Unternehmensleitung die Entscheidung darüber, welche Mitarbeiter befördert werden, dem Zufall überlassen. Man möchte also einen Pseudozufallsvektor $\boldsymbol{x}$ aus E simulieren, wobei alle zulässigen Konfigurationen gleich wahrscheinlich sein sollen. Dies bedeutet, dass man gemäß der Wahrscheinlichkeitsfunktion $\boldsymbol{\pi}$ mit $\pi_{\boldsymbol{x}} = \frac{1}{\ell}$ für jedes $\boldsymbol{x} \in E$ und $\ell = |E|$ simulieren möchte. Insbesondere dann, wenn die Anzahl von Knoten (bzw. Mitarbeitern) und Kanten im Netzwerk groß ist, lässt sich die Anzahl ℓ von zulässigen Konfigurationen nur mit einem großen Aufwand ermitteln, weshalb herkömmliche Simulationsalgorithmen nicht gut anwendbar sind.

Eine Alternative zur Simulation einer zulässigen Konfiguration $\boldsymbol{x}$ aus E stellt der Gibbs-Sampler dar. Es sei $\boldsymbol{q}$ die Wahrscheinlichkeitsfunktion der diskreten Gleichverteilung auf V, d.h., $q_v = \frac{1}{|V|}$ für jedes $v \in V$. Außerdem ist für $v \in V$ und $\boldsymbol{x} \in E$ die bedingte Wahrscheinlichkeitsfunktion $\boldsymbol{\pi}_{\cdot\,|\boldsymbol{x}(-v)}$

gegeben durch

$$\pi_{1|\,\boldsymbol{x}(-v)} = P\big(X(v) = 1 \mid \boldsymbol{X}(-v) = \boldsymbol{x}(-v)\big) = \begin{cases} \frac{1}{2}\,, & \text{wenn } (1, \boldsymbol{x}(-v)) \in E, \\ 0\,, & \text{sonst} \end{cases}$$

und $\pi_{0|\,\boldsymbol{x}(-v)} = 1 - \pi_{1|\,\boldsymbol{x}(-v)}$. Darauf basierend lässt sich der folgende Simulationsalgorithmus zur Erzeugung eines Pseudozufallsvektors $\boldsymbol{x}$ aus E mittels eines Gibbs-Samplers angeben.

1. Wähle als Anfangskonfiguration $\boldsymbol{x}_0$ denjenigen Zustand, bei dem alle Knoten mit 0 markiert sind, d.h., $x_0(v) = 0$ für jedes $v \in V$. Setze $n = 0$.

2. Generiere eine Pseudozufallszahl v gemäß der Wahrscheinlichkeitsfunktion $\boldsymbol{q}$ (d.h., wähle einen Knoten $v \in V$ rein zufällig aus).

3. Wirf eine faire Münze. Wenn die Münze „Kopf" zeigt und $x_n(w) = 0$ für all diejenigen Knoten $w \in V$, die mit v verbunden sind, dann setze $x_{n+1}(v) = 1$, ansonsten setze $x_{n+1}(v) = 0$.

4. Die Werte für alle anderen Knoten $w \neq v$ werden nicht geändert, d.h., $x_{n+1}(w) = x_n(w)$ für jedes $w \in V \setminus \{v\}$. Setze $n = n + 1$.

5. Wenn n groß genug ist, dann akzeptiere $\boldsymbol{x}_n$ als rein zufällig ausgewählte zulässige Konfiguration aus E. Ansonsten geh zurück zu Punkt 2.

6.7.2 Metropolis-Hastings-Algorithmus

Einen weiteren MCMC-Algorithmus stellt der *Metropolis-Hastings-Algorithmus* (nach Nicholas Metropolis, 1915-1999, und W. Keith Hastings, geb. 1930) dar, welcher als Verallgemeinerung des Gibbs-Samplers interpretiert werden kann. Dabei hat zum einen die Übergangsmatrix $\boldsymbol{P}$ der konstruierten Markov-Kette eine allgemeinere Form als in Abschnitt 6.7.1 und zum anderen wird eine Akzeptanz- und Verwerfungsmethode in den Algorithmus integriert.

Wie in Abschnitt 6.7.1 sei V eine endliche Indexmenge und $\boldsymbol{X} = \{X(v), v \in V\}$ ein diskreter Zufallsvektor mit endlichem Wertebereich $E \subset \mathbb{R}^{|V|}$. Des Weiteren sei $\boldsymbol{\pi} = (\pi_{\boldsymbol{x}}, \boldsymbol{x} \in E)$ die Wahrscheinlichkeitsfunktion des Zufallsvektors $\boldsymbol{X}$, sodass $\pi_{\boldsymbol{x}} > 0$ für jeden Zustand $\boldsymbol{x} \in E$. Wir betrachten eine irreduzible und aperiodische stochastische $|E| \times |E|$ Matrix $\boldsymbol{Q} = (q_{\boldsymbol{x}\boldsymbol{x}'})$, d.h., für jedes $\boldsymbol{x} \in E$ bildet $\{q_{\boldsymbol{x}\boldsymbol{x}'}, \boldsymbol{x}' \in E\}$ eine Wahrscheinlichkeitsfunktion auf E, wobei wir annehmen, dass $q_{\boldsymbol{x}'\boldsymbol{x}} > 0$ gilt, wenn $q_{\boldsymbol{x}\boldsymbol{x}'} > 0$ für $\boldsymbol{x}, \boldsymbol{x}' \in E$. Außerdem sei $\boldsymbol{A} = (a_{\boldsymbol{x}\boldsymbol{x}'})$ eine $|E| \times |E|$ Matrix mit $a_{\boldsymbol{x}\boldsymbol{x}'} = \min\left\{1, \frac{\pi_{\boldsymbol{x}'} q_{\boldsymbol{x}'\boldsymbol{x}}}{\pi_{\boldsymbol{x}} q_{\boldsymbol{x}\boldsymbol{x}'}}\right\}$ für jedes Paar $\boldsymbol{x}, \boldsymbol{x}' \in E$.

Satz 6.7.2 Es sei $\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots$ eine Markov-Kette mit Zustandsraum E, beliebiger Anfangsverteilung $\boldsymbol{\alpha}$ und Übergangsmatrix $\boldsymbol{P} = (p_{\boldsymbol{x}\boldsymbol{x}'})$, wobei $p_{\boldsymbol{x}\boldsymbol{x}'} = q_{\boldsymbol{x}\boldsymbol{x}'} a_{\boldsymbol{x}\boldsymbol{x}'}$ für jedes $\boldsymbol{x}, \boldsymbol{x}' \in E$ mit $\boldsymbol{x} \neq \boldsymbol{x}'$ und $p_{\boldsymbol{x}\boldsymbol{x}} = 1 - \sum_{\boldsymbol{x}' \neq \boldsymbol{x}} p_{\boldsymbol{x}\boldsymbol{x}'}$ für $\boldsymbol{x} \in E$. Dann ist $\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots$ irreduzibel und aperiodisch und das Paar $(\boldsymbol{P}, \boldsymbol{\pi})$ ist reversibel. Hieraus folgt insbesondere, dass $\boldsymbol{\pi}$ die eindeutig bestimmte Grenzverteilung der Markov-Kette $\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots$ ist.

Bemerkung 6.7.1

1. Das Theorem besagt, dass der Zufallsvektor $\boldsymbol{X}_n$ für große n näherungsweise gemäß der Wahrscheinlichkeitsfunktion $\boldsymbol{\pi}$ verteilt ist.

2. Die Übergangsmatrix $\boldsymbol{P}$ lässt sich wie folgt interpretieren. Wenn sich die Markov-Kette zum Zeitpunkt n im Zustand $\boldsymbol{x} \in E$ befindet, dann wird zunächst ein möglicher (Folge-) Zustand $\boldsymbol{x}' \in E$ gemäß der Wahrscheinlichkeitsfunktion $\{q_{\boldsymbol{x}\boldsymbol{x}'}, \boldsymbol{x}' \in E\}$ ausgewählt. Dieser wird dann mit Wahrscheinlichkeit $a_{\boldsymbol{x}\boldsymbol{x}'}$ akzeptiert. Wird der vorgeschlagene Zustand abgelehnt, so befindet sich die Markov-Kette zum Zeitpunkt $n+1$ noch immer im Zustand $\boldsymbol{x}$.

3. Ein großer Vorteil des Metropolis-Hastings-Algorithmus ist, dass nur die Quotienten $\frac{\pi_{\boldsymbol{x}'}}{\pi_{\boldsymbol{x}}}$ bekannt sein müssen. Daher eignet sich die Methode insbesondere für Verteilungen, die eine unbekannte Normierungskonstante enthalten (siehe auch das Beispiel am Ende dieses Abschnitts).

Pseudozufallsvektoren mit Wahrscheinlichkeitsfunktion $\boldsymbol{\pi}$ können gemäß dem folgenden Simulationsalgorithmus erzeugt werden.

Simulationsalgorithmus

1. Wähle einen beliebigen (Anfangs-) Zustand $\boldsymbol{x}_0 \in E$ und setze $n = 0$.

2. Generiere einen Pseudozufallsvektor $\boldsymbol{x}'$ gemäß der Wahrscheinlichkeitsfunktion $\{q_{\boldsymbol{x}_n\boldsymbol{x}'}, \boldsymbol{x}' \in E\}$.

3. Erzeuge eine Standard-Pseudozufallszahl u. Wenn $u < \min\left\{1, \frac{\pi_{\boldsymbol{x}'} q_{\boldsymbol{x}'\boldsymbol{x}_n}}{\pi_{\boldsymbol{x}_n} q_{\boldsymbol{x}_n\boldsymbol{x}'}}\right\}$, dann setze $\boldsymbol{x}_{n+1} = \boldsymbol{x}'$. Ansonsten gilt $\boldsymbol{x}_{n+1} = \boldsymbol{x}_n$. Setze $n = n+1$.

4. Wenn n groß genug ist, dann akzeptiere $\boldsymbol{x}_n$ als Pseudozufallsvektor der Wahrscheinlichkeitsfunktion $\boldsymbol{\pi}$. Ansonsten geh zurück zu Punkt 2.

Beispiel 6.7.2 (*Boltzmann-Verteilung*) Der Metropolis-Hastings-Algorithmus kann z.B. zur Simulation von physikalischen Systemen gemäß der Boltzmann-Verteilung (nach Ludwig Eduard Boltzmann, 1844-1906) genutzt werden. Es beschreibe V die (endliche) Struktur des betrachteten Systems und der endliche Zustandsraum $E \subset \mathbb{R}^{|V|}$ beinhalte die Menge aller möglichen (thermodynamischen) Zustände $\boldsymbol{x} = (x(v), v \in V)$, die das physikalische

System annehmen kann. Der Zustand des Systems soll gemäß der Wahrscheinlichkeitsfunktion $\boldsymbol{\pi} = (\pi_{\boldsymbol{x}}, \boldsymbol{x} \in E)$ mit $\pi_{\boldsymbol{x}} = \frac{1}{Z} \exp\left\{-\frac{1}{T} H(\boldsymbol{x})\right\}$ für jedes $\boldsymbol{x} \in E$ simuliert werden. Dabei ist $H(\boldsymbol{x})$ die Energie des Systems im Zustand $\boldsymbol{x}$, T die Temperatur und $Z = \sum_{\boldsymbol{x} \in E} \exp\left\{-\frac{1}{T} H(\boldsymbol{x})\right\}$ eine Normierungskonstante, die gewährleistet, dass $\boldsymbol{\pi}$ eine Wahrscheinlichkeitsfunktion ist. Wenn die Anzahl $|V|$ von Komponenten des Systems groß ist und das physikalische System eine komplizierte Struktur aufweist, sodass die Energie $H(\boldsymbol{x})$ eines Zustands $\boldsymbol{x} \in E$ schwierig zu berechnen ist, dann kann die Konstante Z im Allgemeinen nicht oder nur mit großem Aufwand exakt berechnet werden.

Der Metropolis-Hastings-Algorithmus kann in diesem Fall zur Simulation des Systems verwendet werden. Wir nehmen an, dass sich eine Graphenstruktur auf E bilden lässt. Zum Beispiel können zwei Zustände $\boldsymbol{x}, \boldsymbol{x}' \in E$ als verbunden betrachtet werden, wenn sie in mindestens einer Komponente v identisch sind (d.h. $\boldsymbol{x}(v) = \boldsymbol{x}'(v)$ für mindestens ein $v \in V$) oder wenn sie in mindestens der Hälfte der Komponenten des Systems identisch sind (d.h. $\boldsymbol{x}(v) = \boldsymbol{x}'(v)$ für mindestens die Hälfte aller Komponenten $v \in V$). Eine natürliche Wahl der stochastischen Matrix $\boldsymbol{Q} = (q_{\boldsymbol{x}\boldsymbol{x}'})$ wäre dann durch

$$q_{\boldsymbol{x}\boldsymbol{x}'} = \begin{cases} \dfrac{1}{n_{\boldsymbol{x}}}\,, & \text{wenn } \boldsymbol{x}' \text{ mit } \boldsymbol{x} \text{ verbunden ist,} \\ 0\,, & \text{sonst} \end{cases}$$

gegeben, wobei $n_{\boldsymbol{x}}$ die Anzahl der mit $\boldsymbol{x}$ verbundenen Zustände in E bezeichnet. Um zu gewährleisten, dass $\boldsymbol{Q}$ irreduzibel und aperiodisch ist, nehmen wir an, dass jeder Zustand $\boldsymbol{x} \in E$ mit sich selbst verbunden ist und dass es für je zwei Zustände $\boldsymbol{x}, \boldsymbol{x}' \in E$ eine Folge $\boldsymbol{y}_0, \ldots, \boldsymbol{y}_n \in E$ gibt, sodass $\boldsymbol{x} = \boldsymbol{y}_0, \boldsymbol{x}' = \boldsymbol{y}_n$ und sodass $\boldsymbol{y}_i$ und $\boldsymbol{y}_{i+1}$ verbunden sind für $i = 0, \ldots, n-1$. Die Matrix $\boldsymbol{A} = (a_{\boldsymbol{x}\boldsymbol{x}'})$ der Akzeptanzwahrscheinlichkeiten lässt sich schließlich wie folgt bestimmen:

$$\begin{aligned} a_{\boldsymbol{x}\boldsymbol{x}'} &= \min\left\{1, \frac{\pi_{\boldsymbol{x}'} q_{\boldsymbol{x}'\boldsymbol{x}}}{\pi_{\boldsymbol{x}} q_{\boldsymbol{x}\boldsymbol{x}'}}\right\} = \min\left\{1, \frac{\frac{1}{Z}\exp\left\{-\frac{1}{T}H(\boldsymbol{x}')\right\}\frac{1}{n_{\boldsymbol{x}'}}}{\frac{1}{Z}\exp\left\{-\frac{1}{T}H(\boldsymbol{x})\right\}\frac{1}{n_{\boldsymbol{x}}}}\right\} \\ &= \min\left\{1, \frac{n_{\boldsymbol{x}}}{n_{\boldsymbol{x}'}} \exp\left\{\frac{H(\boldsymbol{x}) - H(\boldsymbol{x}')}{T}\right\}\right\} \end{aligned}$$

für jedes verbundene Paar $\boldsymbol{x}, \boldsymbol{x}' \in E$. Insbesondere muss die Normierungskonstante Z nicht bekannt sein.

Die Wahrscheinlichkeitsfunktion $\boldsymbol{\pi}$ kann dann gemäß dem folgenden Algorithmus simuliert werden.

Simulationsalgorithmus

1. Wähle einen beliebigen (Anfangs-) Zustand $\boldsymbol{x}_0 \in E$ und setze $n = 0$.

2. Wähle rein zufällig einen Zustand $\boldsymbol{x}'$ aus der Menge derjenigen Zustände in E, die mit $\boldsymbol{x}_n$ verbunden sind.

3. Erzeuge eine Standard-Pseudozufallszahl u. Wenn

$$u < \min\left\{1, \frac{n_{\boldsymbol{x}_n}}{n_{\boldsymbol{x}'}} \exp\left\{\frac{H(\boldsymbol{x}_n) - H(\boldsymbol{x}')}{T}\right\}\right\},$$

   dann setze $\boldsymbol{x}_{n+1} = \boldsymbol{x}'$. Ansonsten gilt $\boldsymbol{x}_{n+1} = \boldsymbol{x}_n$. Setze $n = n + 1$.

4. Wenn n groß genug ist, dann akzeptiere $\boldsymbol{x}_n$ als Pseudozufallsvektor der Wahrscheinlichkeitsfunktion $\boldsymbol{\pi}$. Ansonsten geh zurück zu Punkt 2.

Kapitel 7

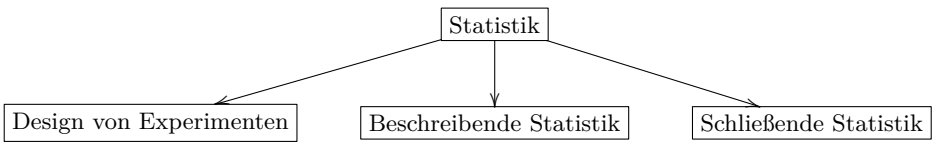
Beschreibende Statistik

7.1 Typische Fragestellungen, Aufgaben und Ziele der Statistik

Im alltaglichen Sprachgebrauch versteht man unter „Statistik“ eine Darstellung von Ergebnissen des Zusammenzählens von Daten und Fakten jeglicher Art, wie z.B. ökonomischen Kenngrößen, politischen Umfragen, Daten der Marktforschung, klinischen Studien in der Biologie und Medizin, usw.

Die *mathematische Statistik* jedoch kann viel mehr. Sie arbeitet mit *Daten-Stichproben,* die nach einem bestimmten Zufallsmechanismus aus der *Grundgesamtheit* aller Daten, die in Folge von Beobachtung, Experimenten (reale Daten) oder Computersimulation (synthetische Daten) erhoben wurden. Dabei beschäftigt sich die mathematische Statistik mit folgenden Fragestellungen:

1. Wie sollen die Daten gewonnen werden? (Design von Experimenten)
2. Wie sollen (insbesondere riesengroße) Datensätze beschrieben werden, um die Gesetzmäßigkeiten und Strukturen in ihnen entdecken zu können? (Beschreibende (deskriptive) und explorative Statistik)
3. Welche Schlüsse kann man aus den Daten ziehen? (Schließende oder induktive Statistik)



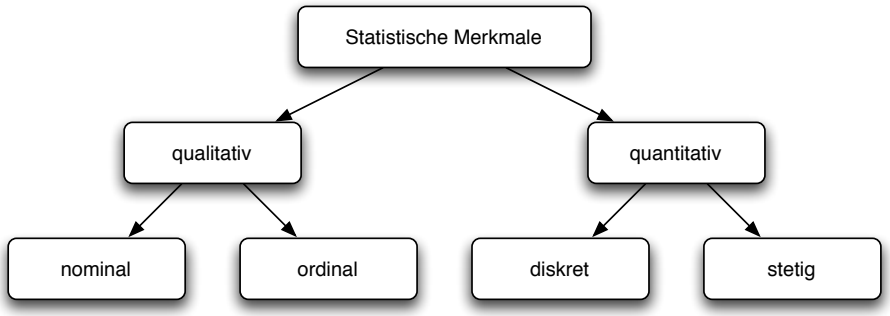
In dieser einführenden Vorlesung werden wir Teile der beschreibenden und schließenden Statistik kennenlernen, wobei die Datenerhebung aus Platzgründen ausgelassen wird. Die *Arbeitsweise eines Statistikers* sieht folgendermaßen aus:

1. *Datenerhebung*
2. *Visualisierung und beschreibende Datenanalyse*
3. *Datenbereinigung* (z.B. Erkennung fehlerhafter Messungen, Ausreißern, usw.)
4. *Explorative Datenanalyse* (Suche nach Gesetzmäßigkeiten)
5. *Modellierung der Daten* mit Methoden der Stochastik
6. *Modellanpassung* (Schätzung der Modellparameter)
7. *Modellvalidierung* (wie gut war die Modellanpassung?)
8. *Schließende Datenanalyse:*
   - Konstruktion von *Vertrauensintervallen* (Konfidenzintervallen) für Modellparameter und deren Funktionen,
   - Tests statistischer Hypothesen,
   - Vorhersage von Zielgrößen (z.B. auf Basis modellbezogener Computersimulation).

Uns werden in diesem Vorlesungsskript vor allem die Arbeitspunkte 2), 4)–6) und 8) beschäftigen.

7.2 Statistische Merkmale und ihre Typen

Die Daten, die zur statistischen Analyse vorliegen, können eine oder mehrere interessierende Größen (die auch *Variablen* oder *Merkmale* genannt werden) umfassen. Ihre Werte werden *Merkmalsausprägungen* genannt. In dem nachfolgenden Diagramm werden mögliche Typen der statistischen Merkmale gegeben.



Diese Typen entstehen in Folge der Klassifikation von Wertebereichen (Skalen) der Merkmale. Dennoch ist diese Einteilung nicht vollständig und kann bei Bedarf erweitert werden. Man unterscheidet *qualitative* und *quantitative* Merkmale. *Quantitative Merkmale* lassen sich inhaltlich gut durch Zahlen darstellen (z.B. Kredithöhe in €, Körpergewicht und Körpergröße, Blutdruck usw.). Sie können *diskrete* oder *stetige* Wertebereiche haben, wobei diskrete Merkmale isolierte Werte annehmen können (z.B. Anzahl der Schäden eines Versicherers pro Jahr). Stetige Wertebereiche hingegen sind überabzählbar. Dennoch liegen in der Praxis stetige Merkmale in gerundeter Form vor (z.B. Körpergröße auf cm gerundet, Geldbetrage auf € gerundet usw.).

Im Gegensatz zu den quantitativen Merkmalen sind die Inhalte der *qualitativen Merkmale*, wie z.B. Blutgruppe (0, A, B und AB) oder Familienstand (ledig, verheiratet, verwitwet), nicht sinnvoll durch Zahlen darzustellen. Sie können zwar formell mit Zahlen kodiert werden (z.B. bei Blutgruppen $0 = 0$, $A = 1$, $B = 2$, $AB = 3$), aber solche Kodierungen stellen keinen inhaltlichen Zusammenhang zwischen Ausprägungen und Zahlen-Codes dar sondern dienen lediglich der besseren Identifikation der Merkmale auf einem Rechner. Es ist insbesondere unsinnig, Mittelwerte und ähnliches von solchen Codes zu bilden.

Ein qualitatives Merkmal mit nur 2 Ausprägungen (z.B. männlich / weiblich, Raucher / Nichtraucher) heißt *alternativ*. Ein qualitatives Merkmal kann *ordinal* (wenn sich eine natürliche lineare Ordnung in den Merkmalsausprägungen finden lässt, wie z.B. gut / mittel / schlecht bei Qualitätsbewertung in Umfragen oder sehr gut / gut / befriedigend / ausreichend / mangelhaft / ungenügend bei Schulnoten) oder *nominal* (wenn eine solche Ordnung nicht vorhanden ist) sein. Beispiele von nominalen Merkmalen sind Fahrzeugmarken in der KFZ-Versicherung (z.B. BMW, Peugeot, Volvo, usw.) oder Führerscheinklassen (A, B, C, ...). Datenmerkmale können auch mehrdimensionale Ausprägungen haben. In dieser Vorlesung behandeln wir jedoch hauptsächlich eindimensionale Merkmale.

7.3 Statistische Daten und Stichproben

Aus den obigen Beispielen wird klar, dass ein Statistiker mit Datensätzen der Form $(x_1, \ldots, x_n)$ arbeitet, wobei die Einzeleinträge x_i aus einer Grundgesamtheit $G \subset \mathbb{R}^k$ stammen, die hypothetisch unendlich groß ist. Der vorliegende Datensatz $(x_1, \ldots, x_n)$ wird auch *(konkrete) Stichprobe* von Umfang n genannt. Die Menge B aller potentiell möglichen Stichproben bezeichnen wir als *Stichprobenraum* und setzen zur Vereinfachung der Notation $B = \mathbb{R}^{kn}$. In diesem Skript werden wir meistens die univariate statistische Analyse (also $k = 1$, ein eindimensionales Merkmal) betreiben. In der beschreibenden Statistik arbeitet man mit Stichproben $(x_1, \ldots, x_n)$ und ihren

Funktionen, um diese Daten visualisieren zu können. Für die Aufgabe der schließenden Statistik jedoch reicht diese Datenebene nicht mehr aus. Daher wird die zweite Ebene der Betrachtung eingeführt, die sogenannte *Modellebene*. Dabei wird angenommen, dass die konkrete Stichprobe $(x_1, \ldots, x_n)$ eine *Realisierung* eines stochastischen Modells $(X_1, \ldots, X_n)$ darstellt, wobei $X_1, \ldots, X_n$ (meistens unabhängige identisch verteilte) Zufallsvariablen auf einem (nicht näher spezifizierten) Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, P)$ sind. Diese Zufallsvariablen $X_i, \quad i = 1, \ldots, n$ können als konsequente Beobachtungen eines Merkmals interpretiert werden.

Der Vektor $(X_1, \ldots, X_n)$ wird dabei *Zufallsstichprobe* genannt. Man setzt weiter voraus, dass $EX_i^2 < \infty \ \forall i = 1, \ldots, n$, damit man von der Varianz Var X_i der Einzeleinträge sprechen kann. Es wird außerdem angenommen, dass ein $\omega \in \Omega$ existiert, sodass $X_i(\omega) = x_i \quad \forall i = 1, \ldots, n$. Sei F die Verteilungsfunktion der Zufallsvariablen X_i. Eine der wichtigsten Aufgaben der Statistik ist die Bestimmung von F (man sagt, „Schätzung von F“) aus den konkreten Daten $(x_1, \ldots, x_n)$. Dabei können auch Momente von F und ihre Funktionen (Erwartungswert, Varianz, Schiefe, usw.) von Interesse sein.

7.4 Stichprobenfunktionen

Um die obigen Aufgaben erfüllen zu können, braucht man gewisse Funktionen $\varphi : \mathbb{R}^n \to \mathbb{R}^m, \quad m \in \mathbb{N}$ auf dem Stichprobenraum, die diese Stichprobe bewerten.

Definition 7.4.1 Eine Borel-messbare Abbildung $\varphi : \mathbb{R}^n \to \mathbb{R}^m$ heißt *Stichprobenfunktion*. Wenn man auf der Modellebene mit einer Zufallsstichprobe $(X_1, \ldots, X_n)$ arbeitet, so heißt die Zufallsvariable

$$\varphi(X_1, \ldots, X_n)$$

eine *Statistik*. In der Schätztheorie spricht man dabei von *Schätzern* und bei statistischen Tests wird $\varphi(X_1, \ldots, X_n)$ *Teststatistik* genannt.

Beispiele für Stichprobenfunktionen sind unter anderen das *Stichprobenmittel*

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i \,,$$

die *Stichprobenvarianz*

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \bar{x}_n\right)^2$$

und die *Ordnungsstatistiken*

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)} \,,$$

die entstehen, wenn man eine Stichprobe, die aus quantitativen Merkmalen besteht, linear ordnet ($x_{(1)} = \min_{i=1,\ldots,n} x_i, \ldots, x_{(n)} = \max_{i=1\ldots n} x_i$). Im Folgenden werden weitere Beispiele und ihre Charakteristiken diskutiert.

Sei eine konkrete Stichprobe $(x_1, \ldots, x_n)\,, \quad x_i \in \mathbb{R}$ gegeben, wobei die x_i als Realisierungen der Zufallsvariablen $X_i \stackrel{d}{=} X$ mit Verteilungsfunktion F interpretiert werden können.

7.5 Verteilungen und ihre Darstellungen

In diesem Abschnitt werden wir Methoden zur statistischen Beschreibung und grafischen Darstellung der (unbekannten) Verteilung F betrachten.

7.5.1 Häufigkeiten und Diagramme

Falls das quantitative Merkmal X eine endliche Anzahl von Ausprägungen $\{a_1, \ldots, a_k\}$, $a_1 < a_2 < \ldots < a_k$, besitzt, also

$$P(X \in \{a_1, \ldots, a_k\}) = 1\,,$$

dann kann eine Schätzung der Zähldichte $p_i = P(X = a_i)$ von X aus den Daten $(x_1, \ldots, x_n)$ grafisch dargestellt werden. Ähnliche Darstellungen sind für die Dichte $f(x)$ von absolut stetigen Merkmalen X möglich, wobei ihr Wertebereich C sich in k Klassen aufteilen lässt: $(c_{i-1}, c_i]\,, \quad i = 1, \ldots, k$, wobei $c_0 = -\infty$, $c_1 < \ldots < c_{k-1}$, $c_k = \infty$ ist. Dann kann die Zähldichte $p_i = P(X \in (c_{k-1}, c_k])$ gegeben durch

$$p_i = \int_{c_{i-1}}^{c_i} f(x)\, dx_i\,, \qquad i = 0, \ldots, k$$

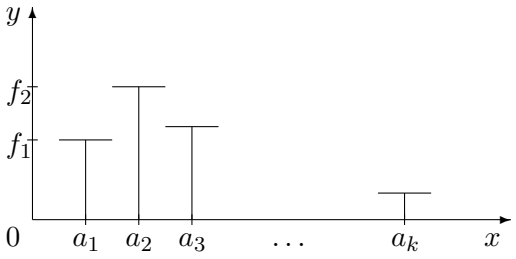
betrachtet werden.

Definition 7.5.1 1. Die *absolute Häufigkeit* von Merkmalsausprägung a_i bzw. Klasse $(c_{i-1}, c_i]\,, \quad i = 1, \ldots, k$ ist $n_i = \#\{x_j,\, j = 1, \ldots, n : x_j = a_i\}$ bzw. $n_i = \#\{x_j,\, j = 1, \ldots, n : x_j \in (c_{i-1}, c_i]\}$.

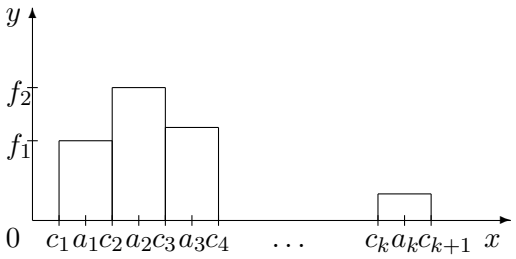
2. Die *relative Häufigkeit* von Merkmalsausprägung a_i bzw. Klasse $(c_{i-1}, c_i]$ ist $f_i = {}^{n_i}\!/_n\,,\ i = 1, \ldots, k$.

Es gilt offensichtlich $n = \sum_{i=1}^k n_i\,, \quad 0 \le f_i \le 1\,, \quad \sum_{i=1}^k f_i = 1$. Die absoluten und relativen Häufigkeiten werden oft in Häufigkeitstabellen zusammengefasst. Zu ihrer Visualisierung dienen so genannte *Diagramme. Histogramme* werden gebildet, indem man die Paare (a_i, f_i) (bzw. $({}^1\!/_2(c_1 + x_{(1)}), f_1)$, $({}^1\!/_2(c_{i-1} + c_i), f_i)$, $i = 2, \ldots, k-1$, $({}^1\!/_2(c_{k-1} + x_{(n)}), f_k)$ im absolut stetigen Fall, wobei hier die Bezeichnung $a_i = {}^1\!/_2(c_{i-1} + c_i)$ verwendet wird und $x_{(1)} < c_1, \quad x_{(n)} > c_{k-1}$ angenommen wird.) auf der Koordinatenebene (x, y) folgendermaßen aufträgt:

- *Stabdiagramm:* f_i wird als Höhe des senkrechten Strichs über a_i dargestellt:



- *Säulendiagramm:* genauso wie ein Stabdiagramm, nur werden Striche durch Säulen der Form $(c_{i-1}, c_i] \times f_i$ ersetzt, wobei im diskreten Fall die Aufteilung der reellen Achse $-\infty = c_0 < c_1 < c_2 < \ldots < c_{k-1} < c_k = \infty$ in Intervalle beliebig vorgenommen werden kann.



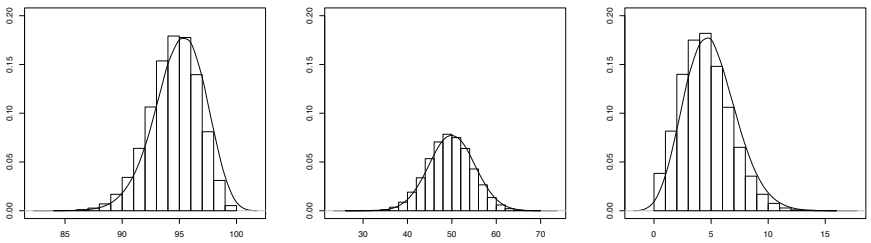


Abbildung 7.1: Das Histogramm der Daten mit einer rechtssteilen (linksschiefen), symmetrischen und linkssteilen (rechtsschiefen) Verteilung und ihre Dichte.

Bemerkung 7.5.1 Die in Abschnitt 7.5.1 betrachteten Methoden dienen der Visualisierung von (Zähl-) Dichten der Verteilung eines beobachteten Merkmals X. Aus dem Histogramm kann z.B. die Interpretation der Form der Dichte abgelesen werden:

Ist die zugrundeliegende Verteilung F_X symmetrisch bzw. linkssteil (rechtsschief) oder rechtssteil (linksschief) (vgl. Abb. 7.1) oder ist sie unimodal (d.h.

eingipflig), bimodal (d.h. mit 2 Gipfeln) oder multimodal (also mit mehreren Gipfeln) (vgl. Abb. 7.2).

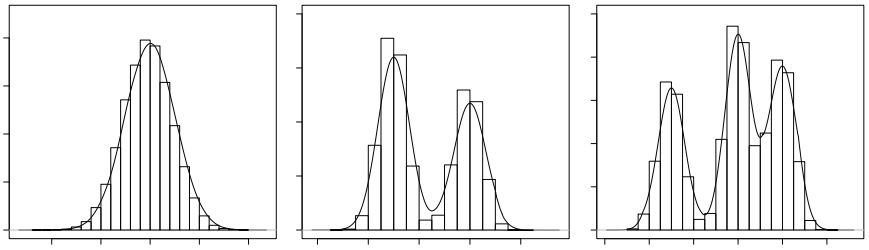


Abbildung 7.2: Histogramm der Daten mit der Dichte einer unimodalen, bimodalen und multimodalen Verteilung

7.5.2 Empirische Verteilungsfunktion

Es sei eine konkrete Stichprobe $(x_1, \ldots, x_n)$ gegeben, die eine Realisierung des statistischen Modells $(X_1, \ldots, X_n)$ ist, wobei $X_1, \ldots, X_n$ unabhängige identisch verteilte Zufallsvariablen mit Verteilungsfunktion $F_X : X_i \stackrel{d}{=} X \sim F_X$ sind. Wie kann die unbekannte Verteilungsfunktion F_X aus den Daten $(x_1, \ldots, x_n)$ rekonstruiert (die Statistiker sagen „geschätzt“) werden? Dies ist mit Hilfe der sogenannten empirischen Verteilungsfunktion möglich:

Definition 7.5.2 1. Die Funktion

$$\hat{F}_n(x) = \#\{x_i : x_i \leq x\,,\, i = 1, \ldots, n\}/n \quad x \in \mathbb{R},$$

heißt *empirische Verteilungsfunktion der konkreten Stichprobe* $x = (x_1, \ldots, x_n)$. Dabei gilt $\hat{F}_n : \mathbb{R}^{n+1} \to [0, 1]$, da $\hat{F}_n(x) = \varphi(x_1, \ldots, x_n, x)$.

2. Die mit $x \in \mathbb{R}$ indizierte Zufallsvariable $\hat{F}_n : \Omega \times \mathbb{R} \to [0, 1]$ heißt *empirische Verteilungsfunktion der Zufallsstichprobe* $(X_1, \ldots, X_n)$, wenn

$$\hat{F}_n(x, \omega) = \hat{F}_n(x) = \frac{1}{n}\#\{X_i,\, i = 1, \ldots, n : X_i(\omega) \leq x\}\,, \quad x \in \mathbb{R}\,.$$

Äquivalent zur Definition 7.5.2 kann man

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq x)\,, \quad x \in \mathbb{R}$$

schreiben, wobei

$$I(x \in A) = \begin{cases} 1, & x \in A \\ 0, & \text{sonst.} \end{cases}$$

Es gilt

$$\hat{F}_n(x) = \begin{cases} 1, & x \geq x_{(n)}\,, \\ \frac{i}{n}, & x_{(i)} \leq x < x_{(i+1)}\,, \quad i = 1, \ldots, n-1\,, \\ 0, & x < x_{(1)}\,. \end{cases}$$

für $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$.

Dabei ist die Höhe des Sprungs an Stelle $x_{(i)}$ gleich der relativen Häufigkeit f_i des Wertes $x_{(i)}$. Falls $x_{(i)} = x_{(i+1)}$ für ein $i \in \{1, \ldots, n\}$, so tritt der Wert i/n nicht auf. In Abbildung 7.3 sieht man, dass $\hat{F}_n(x)$ eine rechtss-

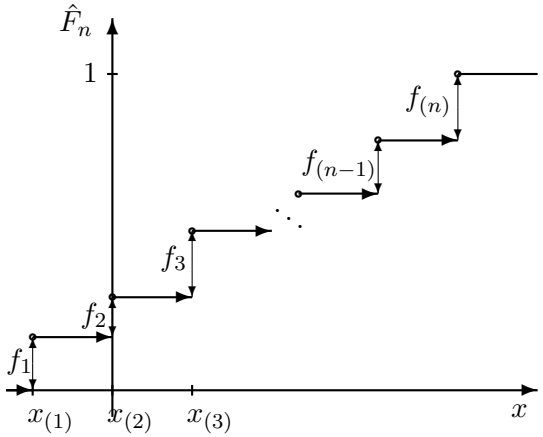
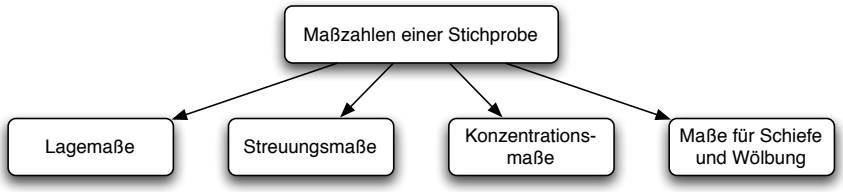


Abbildung 7.3: Eine typische empirische Verteilungsfunktion

tetige monoton nichtfallende Treppenfunktion ist, für die $\hat{F}_n(x) \underset{x \to -\infty}{\longrightarrow} 0$, $\hat{F}_n(x) \underset{x \to \infty}{\longrightarrow} 1$ gilt.

Übungsaufgabe 7.5.1 Zeigen Sie, dass $\hat{F}_n(x)$ eine Verteilungsfunktion ist.

7.6 Beschreibung von Verteilungen



Es sei eine konkrete Stichprobe $(x_1, \ldots, x_n)$ gegeben. Im Folgenden werden Kennzahlen (die sogenannten Maße) dieser Stichprobe betrachtet, welche die wesentlichen Aspekte der der Stichprobe zugrundeliegenden Verteilung wiedergeben:

1. Wo liegen die Werte x_i (Mittel, Ordnungsstatistiken, Quantile)? $\Longrightarrow$ Lagemaße
2. Wie stark streuen die Werte x_i (Varianz) $\Longrightarrow$ Streuungsmaße
3. Wie stark sind die Werte x_i in gewissen Bereichen von $\mathbb{R}$ konzentriert $\Longrightarrow$ Konzentrationsmaße
4. Wie schief bzw. gewölbt ist die Verteilung von $X \Longrightarrow$ Maße für Schiefe und Wölbung

7.6.1 Lagemaße

Man unterscheidet folgende wichtige Lagemaße:

- Mittelwerte: Stichprobenmittel (arithmetisch), geometrisches und harmonisches Mittel, gewichtetes Mittel, getrimmtes Mittel
- Ordnungsstatistiken und Quantile, insbesondere Median und Quartile
- Modus

Betrachten wir sie der Reihe nach:

1. *Mittelwertbildung:* Seit der Antike kennt man mindestens 3 Arten der *Mittelberechnung* von n Zahlen $(x_1, \ldots, x_n)$:
   - *arithmetisch:* $\bar{x}_n = {}^1\!/_n \sum_{i=1}^n x_i\,, \quad \forall x_1, \ldots, x_n \in \mathbb{R},$
   - *geometrisch:* $x_n^g = \sqrt[n]{x_1 \cdot \ldots \cdot x_n}\,, \quad x_1, \ldots, x_n > 0,$
   - *harmonisch:* $x_n^h = \left({}^1\!/_n \sum_{i=1}^n x_i^{-1}\right)^{-1}\,, \quad x_1, \ldots, x_n \neq 0.$

   (a) Das *arithmetische Mittel* wird in der Statistik am meisten benutzt, weil es keine Voraussetzungen über den Wertebereich von $x_1, \ldots, x_n$ braucht. Es wird auch *Stichprobenmittel* genannt. Offensichtlich ist $\bar{x}_n$ ein Spezialfall des sogenannten gewichteten Mittels $x_n^w = \sum_{i=1}^n w_i x_i$, wobei für die Gewichte $w_i \geq 0 \quad \forall i = 1, \ldots, n$ und $\sum_{i=1}^n w_i = 1$ gilt. Als eine natürliche Gewichtewahl kommt $w_i = {}^1\!/_n\,, \quad \forall i = 1, \ldots, n$ bei einer konkreten Stichprobe $(x_1, \ldots, x_n)$ in Frage. Die Summe aller Abweichungen von $\bar{x}_n$ ist Null, denn $\sum_{i=1}^n (x_i - \bar{x}_n) = n\bar{x}_n - n\bar{x}_n = 0$, d.h. $\bar{x}_n$ stellt geometrisch den Schwerpunkt der Werte x_i dar, falls jedem Punkt eine Einheitsmasse zugeordnet wird. Wenn es in der Stichprobe große Ausreißer gibt, so beeinflussen sie das Stichprobenmittel entscheident und erschweren so die objektive Datenanalyse. Deshalb verwendet man oft die robuste Version des arithmetischen

Mittels, das sogenannte *getrimmte Mittel*:

$$\tilde{x}_n^{(k)} = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i)}\,,$$

bei dessen Berechnung die k kleinsten und k größten Ausreißer ausgelassen werden, wobei $k \ll {}^{n}/_{2}$.

(b) Das *geometrische Mittel* wird hauptsächlich bei der Beobachtung von Wachstums- und Zinsfaktoren verwendet. Sei $x_i = {}^{B_i}/_{B_{i-1}}$, $i = 1, \ldots, n$ der Wachstumsfaktor des Merkmals B_i, das in den Jahren $i = 1, \ldots, n$ beobachtet wurde (z.B. Inflationsfaktor). Dann ist $B_n = B_0 \cdot x_1 \cdot \ldots \cdot x_n$ und somit wäre der Zins im Jahre n

$$B_n^g = B_0 \cdot x_1 \cdot \ldots \cdot x_n = B_0 \cdot (x_n^g)^n\ .$$

Für das geometrische Mittel gilt

$$\log x_n^g = \frac{1}{n} \sum_{i=1}^{n} \log x_i \le \log \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)$$

wegen der Konkavität des Logarithmus, d.h. $\log x_n^g = \overline{\log x_n} \le \log \bar{x}_n$ und somit $x_n^g \le \bar{x}_n$, wobei $x_n^g = \bar{x}_n$ genau dann, wenn $x_1 = \ldots = x_n$.

(c) Das *harmonische Mittel* wird bei der Ermittlung von z.B. durchschnittlicher Geschwindigkeiten gebraucht.

Beispiel 7.6.1 Seien x_i Geschwindigkeiten mit denen Bauteile eine Produktionslinie der Länge l durchlaufen. Die gesamte Bearbeitungszeit ist ${}^{l}/_{x_1} + \ldots + {}^{l}/_{x_n}$ und die Durchschnittslaufgeschwindigkeit

$$\frac{l + \ldots + l}{{}^{l}/_{x_1} + \ldots + {}^{l}/_{x_n}} = x_n^h\,.$$

Es gilt $x_{(1)} \le x_n^h \le x_n^g \le \bar{x}_n \le x_{(n)}$ für $x_i > 0$, $i = 1, \ldots, n$.

Übungsaufgabe 7.6.1 Beweisen Sie diese Relation per Induktion bzgl. n.

2. *Ordnungsstatistiken und Quantile*

Definition 7.6.1 Die *Ordnungsstatistiken* $x_{(i)}$, $i = 1, \ldots, n$ der Stichprobe $(x_1, \ldots, x_n)$ sind durch die messbare Permutation $\varphi(x_1, \ldots, x_n)$ gegeben, so dass

$$x_{(i)} = \min \left\{ x_j : \#\{k : x_k \le x_j\} \ge i \right\}, \quad \forall i = 1, \ldots, n\,.$$

Somit gilt $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$. Dieselbe Definition kann auch auf der Modellebene gegeben werden.

Definition 7.6.2 (a) Sei nun X die Zufallsvariable, die das Merkmal modelliert. Sei F_X ihre Verteilungsfunktion. Die verallgemeinerte Inverse von F_X, definiert durch

$$F_X^{-1}(y) = \inf\{x : F_X(x) \geq y\}, \quad y \in [0,1],$$

heißt *Quantilfunktion* von F_X bzw. X. Es gilt $F_X^{-1} : [0,1] \to \mathbb{R} \cup \{\pm\infty\}$. Die Zahl $F_X^{-1}(\alpha)$, $\alpha \in [0,1]$ wird *α-Quantil von F_X* genannt.

(b)
- $F_X^{-1}(0,25)$ heißt *unteres Quartil*,
- $F_X^{-1}(0,75)$ heißt *oberes Quartil*,
- $F_X^{-1}(0,5)$ heißt der *Median* der Verteilung von X.

Zwischen Ordnungsstatistiken und Quantilen besteht ein enger Zusammenhang. So bedeutet $F_X^{-1}(\alpha)$, $\alpha \in (0,1)$, dass ca. $\alpha \cdot 100\%$ aller Merkmalsausprägungen in der Stichprobe $(x_1, \ldots, x_n)$ unter $F_X^{-1}(\alpha)$ und ca. $(1-\alpha) \cdot 100\%$ über $F_X^{-1}(\alpha)$ liegen (im absolut stetigen Fall). Insbesondere gilt $F_X^{-1}(\alpha) \approx x_{([n\alpha])}$, deshalb werden Ordnungsstatistiken auch *empirische Quantile* genannt. Dabei ist x_α definiert als

$$x_\alpha = \begin{cases} x_{([n\alpha]+1)}, & n\alpha \notin \mathbb{N} \\ 1/2(x_{([n\alpha])} + x_{([n\alpha]+1)}), & n\alpha \in \mathbb{N} \end{cases}.$$

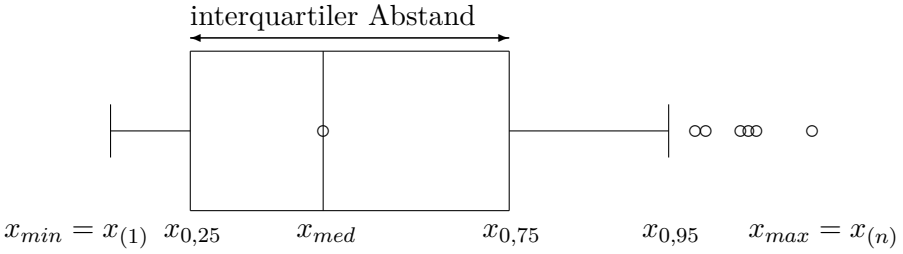
Dies ist die allgemeine Definition des *empirischen α-Quantils.*

Der *empirische Median* ist

$$x_{med} = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & n \text{ ungerade} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right), & n \text{ gerade.} \end{cases}$$

Somit sind mindestens 50% aller Stichprobenwerte kleiner gleich und 50% größer gleich x_{med}. Der Median ist ein Lagemaß, das ein robuster Ersatz für den Mittelwert darstellt, denn er ist bzgl. Ausreißern in der Stichprobe nicht sensibel.

Die oben genannten Statistiken werden in einem *Box-Plot* zusammengefasst und grafisch dargestellt:



Manchmal werden $x_{(1)}$ und $x_{(n)}$ durch $x_{0,05}$ und $x_{0,95}$ ersetzt. Die restlichen Werte werden darüber hinaus als Einzelpunkte auf der x-Achse abgebildet. Dann liegt ein sogenannter *modifizierter Box-Plot* vor.

3. *Modus:* Sei $(x_1, \ldots, x_n)$ eine Stichprobe, die aus n unabhängigen Realisierungen des Merkmals X besteht. Sei $(p(x))$ $f(x)$ die (Zähl-) Dichte von X, wobei die Verteilung von X undimodal ist.

   Definition 7.6.3 (a) Der Wert $x_{mod} = \arg\max f(x)$ $(\arg\max p(x))$ wird der *Modus der Verteilung von X* genannt (vgl. Abb. 7.4).

   (b) Empirisch wird $\hat{x}_{mod}$ als $\frac{c_{m-1}+c_m}{2}$ für $m = \arg\max f_i$ definiert, also als die Mitte des Intervalls mit der größten Häufigkeit des Vorkommens in der Stichprobe, falls dieser eindeutig bestimmbar ist.

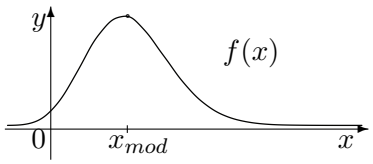


Abbildung 7.4: Veranschaulichung des Modus

Den Mittelwert $\bar{x}_n$, Median x_{med} und Modus x_{mod} kann man auch wie folgt definieren:

$$\bar{x}_n = \arg\min_{x \in \mathbb{R}} \sum_{i=1}^{n} (x_i - x)^2$$

$$x_{med} = \arg\min_{x \in \mathbb{R}} \sum_{i=1}^{n} |x_i - x|$$

$$\hat{x}_{mod} = \frac{c_{m-1}+c_m}{2}, \qquad \text{wobei } m = \arg\min_{j=1,\ldots,n} \sum_{i=1}^{n} I(x_i \notin (c_{j-1}, c_j])$$

Übungsaufgabe 7.6.2 Zeigen Sie die Äquivalenz der oben genannten Definitionen des Mittelwerts $\bar{x}_n$, Medians x_{med} und des Modus x_{mod} zu den bekannten Definitionen.

Die Größen $\bar{x}_n$, x_{med} und $\hat{x}_{mod}$ können auch zur Beschreibung der Symmetrie einer unimodalen Verteilung F_X von Daten $(x_1, \ldots, x_n)$ verwendet werden, da

- bei symmetrischen Verteilung F_X gilt $\bar{x}_n \approx x_{med} \approx \hat{x}_{mod}$
- bei linkssteilen Verteilung F_X gilt $\hat{x}_{mod} < x_{med} < \bar{x}_n$
- bei rechtssteilen Verteilung F_X gilt $\bar{x}_n < x_{med} < \hat{x}_{mod}$.

7.6.2 Streuungsmaße

Bekannte Streuungsmaße einer konkreten Stichprobe $(x_1, \ldots, x_n)$ sind die folgenden Größen:

- *Spannweite* $x_{(n)} - x_{(1)}$,
- *empirische Varianz* $\bar{s}_n^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}_n)^2$,
- *Stichprobenvarianz* $s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x}_n)^2 = \frac{n}{n-1}\bar{s}_n^2$,
- *empirische Standardabweichungen* $\bar{s}_n = \sqrt{\bar{s}_n^2}$, $s_n = \sqrt{s_n^2}$,
- *empirischer Variationskoeffizient* $\gamma_n = {}^{s_n}/_{\bar{x}_n}$, falls $\bar{x}_n > 0$.

Die Spannweite zeigt die *maximale Streuung* in den Daten, wobei sich die empirische Varianz mit der *mittleren quadratischen Abweichung* vom Stichprobenmittel auseinandersetzt. Hier sind einige Eigenschaften von $\bar{s}_n^2$ (bzw. s_n^2, da sie sich nur durch einen Faktor unterscheiden):

Lemma 7.6.1 1. Für jedes $b \in \mathbb{R}$ gilt

$$\sum_{i=1}^n (x_i - b)^2 = \sum_{i=1}^n (x_i - \bar{x}_n)^2 + n(\bar{x}_n - b)^2$$

und somit für $b = 0$

$$\bar{s}_n^2 = \frac{1}{n}\sum_{i=1}^n \left(x_i^2 - \bar{x}_n^2\right) \qquad \text{bzw.} \qquad s_n^2 = \frac{1}{n-1}\sum_{i=1}^n \left(x_i^2 - \bar{x}_n^2\right).$$

2. *Transformationsregel:*
Falls die Daten $(x_1, \ldots, x_n)$ linear transformiert werden, d.h. $y_i = ax_i + b$, $a \neq 0$, $b \in \mathbb{R}$, dann gilt

$$\bar{s}_{n,y}^2 = a^2 \bar{s}_{n,x}^2 \qquad \text{bzw.} \qquad \bar{s}_{n,y} = |a|\bar{s}_{n,x},$$

wobei

$$\bar{s}_{n,y}^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y}_n)^2, \qquad \bar{s}_{n,x}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}_n)^2$$

Beweis 1. Es gilt:

$$\begin{aligned}
&\sum_{i=1}^{n}(x_i-b)^2=\sum_{i=1}^{n}(x_i-\bar{x}_n+\bar{x}_n-b)^2\\
&=\sum_{i=1}^{n}(x_i-\bar{x}_n)^2+2\sum_{i=1}^{n}(x_i-\bar{x}_n)\cdot(\bar{x}_n-b)+\sum_{i=1}^{n}(\bar{x}_n-b)^2\\
&=\sum_{i=1}^{n}(x_i-\bar{x}_n)^2+2(\bar{x}_n-b)\cdot\underbrace{\sum_{i=1}^{n}(x_i-\bar{x}_n)}_{=0}+n(\bar{x}_n-b)^2\,,\quad\forall b\in\mathbb{R}\,.
\end{aligned}$$

2. Es gilt:

$$\bar{s}^2_{n,y}=\frac{1}{n}\sum_{i=1}^{n}(ax_i+b-a\bar{x}_n-b)^2=\frac{a^2}{n}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2=a^2\bar{s}^2_{n,x}\,.$$

□

Der Skalierungsunterschied zwischen $\bar{s}_n^2$ und s_n^2 ist den Eigenschaften der *Erwartungstreue* von s_n^2 zu verdanken, die später im Laufe dieser Vorlesung behandelt wird, und besagt, dass für eine Zufallsstichprobe $(X_1,\ldots,X_n)$ mit X_i unabhängig identisch verteilt, $X_i\sim X$, Var $X=\sigma^2\in(0,\infty)$ gilt $\mathrm{E}s_n^2=\sigma^2$, wobei $\mathrm{E}\bar{s}_n^2=\frac{n}{n-1}\sigma^2\underset{n\to\infty}{\longrightarrow}\sigma^2$. Das heißt, während bei der Verwendung von s_n^2 zur Schätzung von σ^2 kein Fehler „im Mittel“ gemacht wird, ist diese Aussage für $\bar{s}_n^2$ nur asymptotisch (für große Datenmengen n) richtig.

Aufgrund von $\sum_{i=1}^n(x_i-\bar{x}_n)=0$ ist z.B. $x_n-\bar{x}_n$ durch $x_i-\bar{x}_n$, $i=1,\ldots,n-1$ bestimmt. Somit verringert sich die *Anzahl der Freiheitsgrade* in der Summe $\sum_{i=1}^n(x_i-\bar{x}_n)^2$ um 1 und somit scheint die Normierung $\frac{1}{n-1}$ plausibel zu sein.

Die *Standardabweichungen* $\bar{s}_n$ und s_n werden verwendet, damit man die selben Einheiten (und nicht ihre Quadrate, also z.B. Euro und nicht Euro^2) erhält. Für normalverteilte Stichproben ($X\sim N(\mu,\sigma^2)$) liefert $\bar{s}_n$ auch die „k-Sigma-Regel“ (vgl. Vorlesung WR), die besagt, dass in den Intervallen

$$\begin{aligned}
[\bar{x}_n-\bar{s}_n,\bar{x}_n+\bar{s}_n] &\quad\text{ca.}\quad 68\%\,,\\
[\bar{x}_n-2\bar{s}_n,\bar{x}_n+2\bar{s}_n] &\quad\text{ca.}\quad 95\%\,,\\
[\bar{x}_n-3\bar{s}_n,\bar{x}_n+3\bar{s}_n] &\quad\text{ca.}\quad 99\%
\end{aligned}$$

aller Daten liegen.

Der Vorteil vom *empirischen Variationskoeffizienten* ist, dass er *maßstabsunabhängig* ist und somit den Vergleich von Streuungseigenschaften unterschiedlicher Stichproben zulässt.

7.6.3 Maße für Schiefe und Wölbung

Im Vorlesungsskript WR, Abschnitt 4.5 S. 99 wurden folgende Maße für Schiefe bzw. Wölbung der Verteilung einer Zufallsvariable X eingeführt:
Schiefe oder Symmetriekoeffizient:

$$\gamma_1 = \frac{\mu_3'}{\sigma^3} = E(\tilde{X}^3)\,,$$

wobei

$$\mu_k' = \mathrm{E}(X - \mathrm{E}X)^k\,, \quad \sigma^2 = \mu_2' = \mathrm{Var}\ X\,, \quad \tilde{X} = \frac{X - \mathrm{E}X}{\sigma}\,.$$

Wölbung (Exzess):

$$\gamma_2 = \frac{\mu_4'}{\sigma^4} - 3 = \mathrm{E}(\tilde{X}^4) - 3\,,$$

vorausgesetzt, dass $\mathrm{E}(X^4) < \infty$. Für ihre Bedeutung und Interpretation siehe die oben genannten Seiten des WR-Vorlesungsskriptes. Falls nun das Merkmal X statistisch in einer Stichprobe $(x_1, \ldots, x_n)$ beobachtet wird, wie können γ_1 und γ_2 aus diesen Daten geschätzt und interpretiert werden?

Als Schätzer für das k-te zentrierte Moment $\mu_k' = \mathrm{E}(X - \mathrm{E}X)^k$, $k \in \mathbb{N}$ schlagen wir

$$\hat{\mu}_k' = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^k$$

vor, die Varianz σ^2 wird durch

$$\bar{s}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$$

geschätzt. Somit bekommt man den Momentenkoeffizient an der Schiefe (engl. „skewness“)

$$\hat{\gamma}_1 = \frac{\hat{\mu}_3'}{\bar{s}_n^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2\right)^{3/2}}\,.$$

Falls die Verteilung von X linksschief ist, überwiegen positive Abweichungen im Zähler und somit gilt $\hat{\gamma}_1 > 0$ für linksschiefe Verteilungen. Analog gilt $\hat{\gamma}_1 \approx 0$ für symmetrische und $\hat{\gamma}_1 < 0$ für rechtsschiefe Verteilungen.

Das *Wölbungsmaß von Fisher* (engl. „kurtosis“) ist gegeben durch

$$\hat{\gamma}_2 = \frac{\hat{\mu}_4'}{\bar{s}_n^4} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n^2)\right)^2} - 3\,.$$

Falls $\hat{\gamma}_2 > 0$ so ist die Verteilung von X steilgipflig, für $\hat{\gamma}_2 < 0$ ist sie flachgipflig. Falls $X \sim N(\mu, \sigma^2)$, so gilt $\hat{\gamma}_2 \approx 0$. Die Ursache dafür ist, dass

die steilgipfligen Verteilungen schwerere Tails haben als die flachgipfligen. Als Maß dient dabei die Normalverteilung, für die $\gamma_1 = \gamma_2 = 0$ und somit $\hat{\gamma}_1 \approx 0$, $\hat{\gamma}_2 \approx 0$. So definiert, sind $\hat{\gamma}_1$ und $\hat{\gamma}_2$ nicht resistent gegenüber Ausreissern. Eine robuste Variante von $\hat{\gamma}_1$ ist beispielsweise durch den sogennanten *Quantilskoeffizienten der Schiefe*

$$\hat{\gamma}_q(\alpha) = \frac{(x_{1-\alpha} - x_{med}) - (x_{med} - x_\alpha)}{x_{1-\alpha} - x_\alpha}, \quad \alpha \in (0, 1/2)$$

gegeben.

Für $\alpha = 0,25$ erhält man den Quartilskoeffizienten. $\hat{\gamma}_q(\alpha)$ misst den Unterschied zwischen der Entfernung des α- und $(1-\alpha)$-Quantils zum Median. Bei linkssteilen (bzw. rechtssteilen) Verteilungen liegt das (untere) x_α-Quantil näher an (bzw. weiter entfernt von) dem Median. Somit gilt

- $\hat{\gamma}_q(\alpha) > 0$ für linkssteile Verteilungen,
- $\hat{\gamma}_q(\alpha) < 0$ für rechtssteile Verteilungen,
- $\hat{\gamma}_q(\alpha) = 0$ für symmetrische Verteilungen.

Durch das zusätzliche Normieren (Nenner) gilt $-1 \leq \hat{\gamma}_q(\alpha) \leq 1$.

7.7 Quantilplots (Quantil-Grafiken)

Nach der ersten beschreibenden Analyse eines Datensatzes $(x_1, \ldots, x_n)$ soll überlegt werden, mit welcher Verteilung diese Stichprobe modelliert werden kann. Hier sind die sogenannten *Quantilplots* behilflich, da sie grafisch zeigen, wie gut die Daten $(x_1, \ldots, x_n)$ mit dem Verteilungsgesetz G übereinstimmen, wobei G die Verteilungsfunktion einer hypothetischen Verteilung ist.

Sei X eine Zufallsvariable mit (unbekannter) Verteilungsfunktion F_X. Auf Basis der Daten $(X_1, \ldots, X_n)$, X_i unabhängig identisch verteilt und $X_i \stackrel{d}{=} X$ möchte man prüfen, ob $F_X = G$ für eine bekannte Verteilungsfunktion G gilt. Die Methode der *Quantil-Grafiken* besteht darin, dass man die entsprechenden Quantil-Funktionen $\hat{F}_n^{-1}$ und G^{-1} von $\hat{F}_n$ und G grafisch vergleicht. Hierzu

- plotte man $G^{-1}(k/n)$ gegen $\hat{F}_n^{-1}(k/n) = X_{(k)}$, $\quad k = 1, \ldots, n$.
- Falls die Punktwolke

$$\left\{ \left(G^{-1}(k/n),\, X_{(k)} \right), \quad k = 1, \ldots, n \right\}$$

näherungsweise auf einer Geraden $y = ax + b$ liegt, so sagt man, dass $F_X(x) \approx G\left(\frac{x-a}{b}\right)$, $x \in \mathbb{R}$.

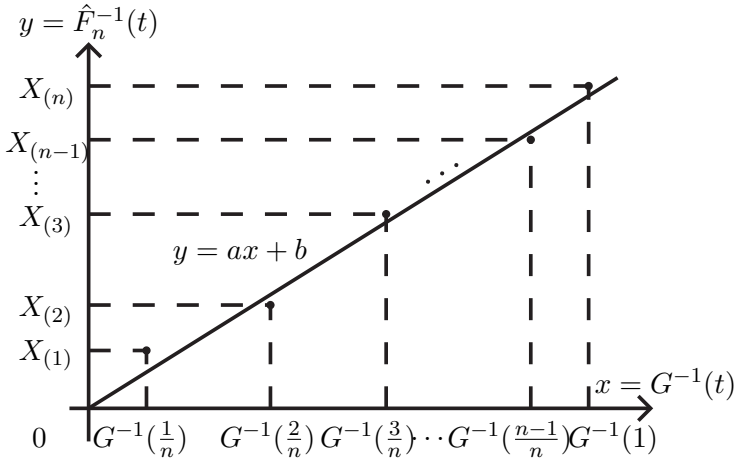


Abbildung 7.5: Quantil-Grafik

Diese empirische Vergleichsmethode beruht auf folgenden Überlegungen:

- Man ersetzt die unbekannte Funktion F_X durch die aus den Daten berechenbare Funktion $\hat{F}_n$. Dabei macht man einen Fehler, der allerdings asymptotisch (für $n \to \infty$) klein ist. Dies folgt aus dem Satz 8.3.9 (WR-Skript) von Gliwenko-Cantelli, der besagt, dass

$$\sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F_X(x) \right| \underset{n \to \infty}{\longrightarrow} 0 \,.$$

  Der Vergleich der entsprechenden Quantil-Funktionen wird durch folgendes Ergebnis bestärkt: Falls $\mathrm{E}X < \infty$, dann gilt

$$\sup_{t \in [0,1]} \left| \int_0^t \left(\hat{F}_n^{-1}(y) - F_X^{-1}(y) \right) dy \right| \underset{n \to \infty}{\overset{\text{f.s.}}{\longrightarrow}} 0 \,.$$

  Somit setzt man bei der Verwendung der Quantil-Grafiken voraus, dass der Stichprobenumfang n ausreichend groß ist, um $\hat{F}_n^{-1} \approx F_X^{-1}$ zu gewährleisten.

- Man setzt zusätzlich voraus, dass die Gleichungen

$$\begin{aligned} y &= ax + b \,, \\ y &= F_X^{-1}(t) \,, \\ x &= G^{-1}(t) \end{aligned}$$

  für alle t (und nicht nur näherungsweise für $t = {}^{k}\!/_{n}$, $k = 1, \ldots, n$) gelten. Daraus folgt, dass $G(x) = t = F_X(y) = F_X(ax + b)$ für alle x, oder $F_X(y) = G\left(\frac{y-b}{a}\right)$ für alle y, weil $x = \frac{y-b}{a}$ ist.

Aus praktischer Sicht ist es besser, Paare $\left(G^{-1}\left(\frac{k}{n+1}\right), X_{(k)}\right), \quad k = 1, \ldots, n$ zu plotten. Dadurch wird vermieden, dass $G^{-1}(n/n) = G^{-1}(1) = \infty$ vorkommt, wie es zum Beispiel bei einer Verteilung G der Fall ist, bei der $F(x) < 1$ gilt für alle $x \in \mathbb{R}$. Tatsächlich gilt für $k = n$, dass $\frac{n}{n+1} < 1$ und somit $G^{-1}\left(\frac{n}{n+1}\right) < \infty$.

Beispiel 7.7.1 (Exponential-Verteilung, $G(x) = (1 - e^{-\lambda x}) \cdot I(x \geq 0)$) Es gilt $G^{-1}(y) = -1/\lambda \log(1 - y), \quad y \in (0, 1)$. So wird man beim Quantil-Plot Paare

$$\left(-\frac{1}{\lambda} \log\left(1 - \frac{k}{n+1}\right), X_{(k)}\right), \quad k = 1, \ldots, n$$

zeichnen, wobei der Faktor $1/\lambda$ für die Linearität unwesentlich ist und weggelassen werden kann.

Beispiel 7.7.2 (Normalverteilung, $G(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt, x \in \mathbb{R}$) Leider ist die analytische Berechnung von Φ^{-1} mit einer geschlossenen Formel nicht möglich. Aus diesem Grund wird $\Phi^{-1}\left(\frac{k}{n+1}\right)$ numerisch berechnet und in Tabellen oder statistischen Software-Paketen (wie z.B. R) abgelegt. Um die empirische Verteilung der Daten mit der Normalverteilung zu vergleichen, trägt man Punkte mit Koordinaten

$$\left(\Phi^{-1}\left(\frac{k}{n+1}\right), X_{(k)}\right), \quad k = 1, \ldots, n$$

auf der Ebene auf und prüft, ob sie eine Gerade bilden (vgl. Abb. 7.6).

Bemerkung 7.7.1 Falls $\bar{x}_n = 0$ und die Verteilung F_X linkssteil ist, so sind die Quantile von F_X kleiner als die von Φ. Somit ist der Normal-Quantilplot konvex. Falls $\bar{x}_n = 0$ und F_X rechtssteil ist, so wird der Normal-Quantilplot konkav sein.

Beispiel 7.7.3 (Haftpflichtversicherung (Belgien, 1992)) In Abbildung 7.7 sind Ordnungsstatistiken der Stichprobe von $n = 227$ Schadenhöhen der Industrie-Unfälle in Belgien im Jahr 1992 (Haftpflichtversicherung) gegen die Quantile von Exponential-, Pareto-, Standardnormal- sowie Weibull-Verteilungen geplottet. Im Bereich von Kleinschäden zeigen die Exponential- und Pareto-Verteilungen eine gute Übereinstimmung mit den Daten. Die Verteilung von mittelgroßen Schäden kann am besten durch die Lognormal- und Weibul-Verteilungen modelliert werden. Für Großschäden erweist sich die Weibull-Verteilung als geeignet.

Beispiel 7.7.4 (Rendite der BMW-Aktie) In Abbildung 7.8 ist der Quantilplot für Renditen der BMW-Aktie beispielhaft zu sehen.

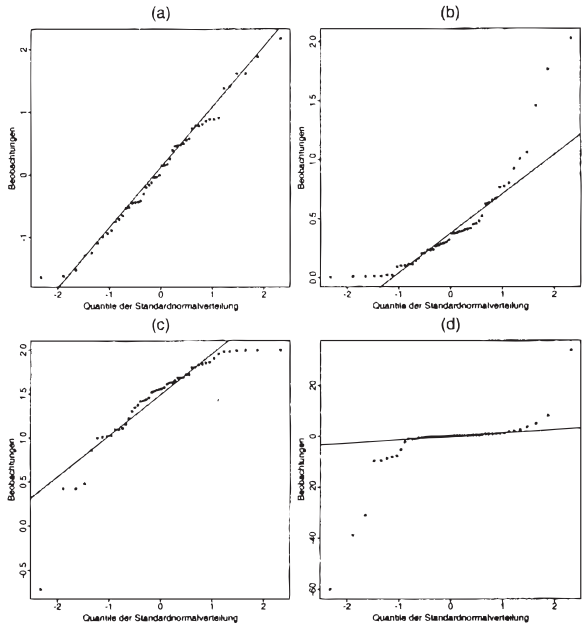


Abbildung 7.6: QQ-Plot einer Normalverteilung (a), einer linkssteilen Verteilung (b), einer rechtssteilen Verteilung (c) und einer symmetrischen, aber stark gekrümmten Verteilung (d)

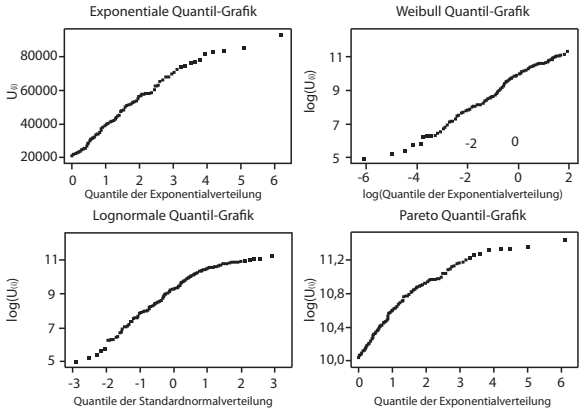


Abbildung 7.7: Ordnungsstatistiken einer Stichprobe von Schadenhöhen der Industrie-Unfälle in Belgien im Jahr 1992

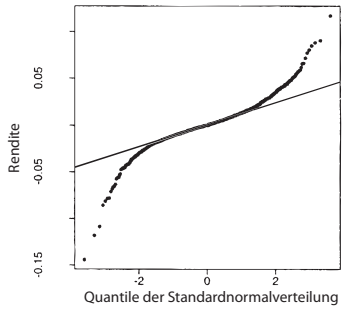


Abbildung 7.8: Quantilplot der Rendite der BMW-Aktie

7.8 Dichteschätzung

Sei eine Stichprobe $(x_1, \ldots, x_n)$ von unabhängigen Realisierungen eines absolut stetig verteilten Merkmals X mit Dichte f_X gegeben. Mit Hilfe der in Abschnitt 7.5.1 eingeführten Histogramme lässt sich f_X grafisch durch eine Treppenfunktion $\hat{f}_X$ darstellen. Dabei gibt es zwei entscheidende Nachteile der Histogrammdarstellung:

1. Willkür in der Wahl der Klasseneinteilung $[c_{i-1}, c_i]$,
2. Eine (möglicherweise) stetige Funktion f_X wird durch eine Treppenfunktion $\hat{f}_X$ ersetzt.

In diesem Abschnitt werden wir versuchen, diese Nachteile zu beseitigen, indem wir eine Klasse von Kerndichtenschätzern einführen, die (je nach Wahl des Kerns) auch zu stetigen Schätzern $\hat{f}_X$ führen.

Definition 7.8.1 Der Kern $K(x)$ wird definiert als eine nicht-negative messbare Funktion auf $\mathbb{R}$ mit der Eigenschaft $\int_{\mathbb{R}} K(x)\, dx = 1$.

Definition 7.8.2 Der *Kerndichteschätzer* der Dichte f_X aus den Daten $(x_1, \ldots, x_n)$ mit Kernfunktion $K(x)$ ist gegeben durch

$$\hat{f}_X(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \quad x \in \mathbb{R},$$

wobei $h > 0$ die sogenannte *Bandbreite* ist.

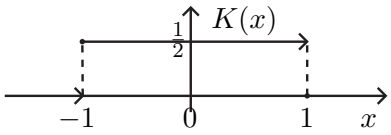
Beispiele für Kerne:

1. *Rechteckskern:*
$$K(x) = 1/2 \cdot I(x \in [-1, 1)).$$
Dabei ist
$$\frac{1}{h} K\left(\frac{x - x_i}{h}\right) = \begin{cases} 1/(2h), & x_i - h \leq x < x_i + h, \\ 0, & \text{sonst}, \end{cases}$$

und somit

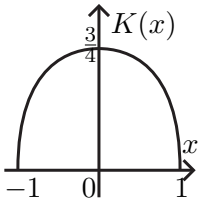
$$\hat{f}_X(x) = \frac{1}{nh} \sum_{i=1}^{k} K\left(\frac{x - x_i}{h}\right) = \frac{\#\{x_i \in [x-h, x+h)\}}{2nh},$$

das auch *gleitendes Histogramm* genannt wird. Dieser Dichteschätzer ist (noch) nicht stetig, was durch die (besonders einfache rechteckige unstetige) Form des Kerns erklärt wird.



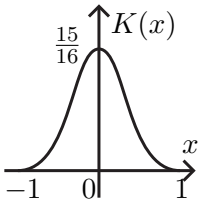
2. *Epanechnikov-Kern:*

$$K(x) = \begin{cases} 3/4(1 - x^2), & x \in [-1, 1) \\ 0, & \text{sonst.} \end{cases}$$



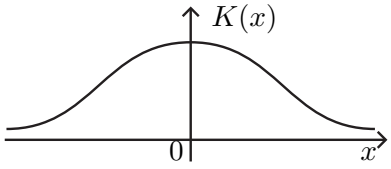
3. *Bisquare-Kern:*

$$K(x) = \frac{15}{16}\left((1 - x^2)^2 \cdot I(x \in [-1, 1))\right).$$



4. *Gauss-Kern:*

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \mathbb{R}.$$



Dabei ist die Wahl der Bandbreite h entscheidend für die Qualität der Schätzung. Je größer $h > 0$, desto glatter wird $\hat{f}_X$ sein und desto mehr „Details“ werden „herausgemittelt“. Für kleinere h wird $\hat{f}_X$ rauer. Dabei können aber auch Details auftreten, die rein stochastischer Natur sind und keine Gesetzmäßigkeiten zeigen. Mit der adäquaten Wahl von h beschäftigen sich viele wissenschaftliche Arbeiten, die empirische Faustregeln, aber auch kompliziertere Optimierungsmethoden dafür vorschlagen. Insgesamt ist das Problem der optimalen Dichteschätzung in der Statistik immer noch offen.

7.9 Beschreibung und Exploration von bivariaten Datensätzen

Im Gegensatz zu der Datenlage in den Abschnitten 7.5 bis 7.8 betrachten wir im Folgenden Datensätze bestehend aus 2 Stichproben $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$, die als Realisierungen von stochastischen Stichproben (Zufallsvektoren) $(X_1, \ldots, X_n)$ und $(Y_1, \ldots, Y_n)$ aufgefasst werden, wobei $X_1, \ldots, X_n$ unabhängige identisch verteilte Zufallsvariablen mit $X_i \stackrel{d}{=} X \sim F_X$ und $Y_1, \ldots, Y_n$ unabhängige identisch verteilte Zufallsvariablen mit $Y_i \stackrel{d}{=} Y \sim F_Y$ sind. Wir betrachten hier ausschließlich quantitative Merkmale X und Y. Es wird ein Zusammenhang zwischen X und Y vermutet, der an Hand von (konkreten) Stichproben $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$ näher untersucht werden soll. Mit anderen Worten, wir interessieren uns für die Eigenschaften der bivariaten Verteilung $F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$ des Zufallsvektors $(X, Y)^T$.

7.9.1 Zusammenhangsmaße

Jetzt wird uns die Frage beschäftigen, in welchem Maße die Merkmale X und Y voneinander abhängig sind. Um die $\mathrm{Cov}\,(X, Y) = \mathrm{E}(X - \mathrm{E}X)(Y - \mathrm{E}Y)$ aus den Daten zu schätzen, setzt man die sogenannte *empirische Kovarianz*

$$S_{xy}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)(y_i - \bar{y}_n)$$

ein. Dabei ist S_{xy}^2 jedoch von den Skalen von X und Y abhängig.

1. Um eine skaleninvariantes Zusammenhangsmaß zu bekommen, betrachtet man die empirische Variante des Korrelationskoeffizienten

$$\varrho(X, Y) = \frac{\mathrm{Cov}\,(X, Y)}{\sqrt{\mathrm{Var}\;X} \cdot \sqrt{\mathrm{Var}\;Y}}\,,$$

den sogenannten *Bravais-Pearson-Korrelationskoeffizienten*

$$\varrho_{xy} = \frac{S_{xy}^2}{\sqrt{S_{xx}^2 \cdot S_{yy}^2}}\,,$$

wobei

$$S_{xx}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2\,, \qquad S_{yy}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2$$

die Stichprobenvarianzen der Stichproben $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$ sind. Dabei erbt ϱ_{xy} alle Eigenschaften des Korrelationskoeffizienten $\varrho(X, Y)$:

(a) $|\varrho_{xy}| \leq 1$

(b) $\varrho_{xy} = \pm 1$, falls ein linearer Zusammenhang in den Stichproben $(x_i, y_i)_{i=1,\ldots,n}$ vorliegt, d.h. alle Punkte (x_i, y_i), $i = 1, \ldots, n$ liegen auf einer Gerade mit positivem (bei $\varrho_{xy} = 1$) bzw. negativem (bei $\varrho_{xy} = -1$) Anstieg.

(c) Wenn $|\varrho_{xy}|$ klein ist ($\varrho_{xy} \approx 0$), so sind die Datensätze unkorreliert. Dabei wird oft folgende grobe Einteilung vorgenommen:

Merkmale X und Y sind

- *„schwach korreliert“*, falls $|\varrho_{xy}| < 0.5$,
- *„stark korreliert“*, falls $|\varrho_{xy}| \geq 0.8$.

Ansonsten liegt ein mittlerer Zusammenhang zwischen X und Y vor.

Lemma 7.9.1 Für ϱ_{xy} gilt die alternative rechengünstige Darstellung

$$\varrho_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}_n\bar{y}_n}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}_n^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}_n^2\right)}}\,. \tag{7.1}$$

2. *Spearmans Korrelationskoeffizient*
Einen alternativen Korrelationskoeffizienten erhält man, wenn man die Stichprobenwerte x_i bzw. y_i in ϱ_{xy} durch ihre *Ränge* $\mathrm{rg}(x_i)$ bzw. $\mathrm{rg}(y_i)$ ersetzt, die als Position dieser Werte in den ansteigend geordneten Stichproben zu verstehen sind:
$\mathrm{rg}(x_i) = j$, falls $x_i = x_{(j)}$ für ein $j \in \{1, \ldots, n\}$, $\forall i = 1, \ldots, n$. Es bedeutet, dass $\mathrm{rg}(x_{(i)}) = i\ \forall i = 1, \ldots, n$, falls $x_i \neq x_j$ für $i \neq j$.

Falls die Stichprobe $(x_1, \ldots, x_n)$ k identische Werte x_i (die sogenannten *Bindungen*) enthält, so wird diesen Werten der sogenannte Durchschnittsrang $\mathrm{rg}(x_i)$ zugewiesen, der als arithmetisches Mittel der k in Frage kommenden Ränge errechnet wird. Zum Beispiel findet folgende Zuordnung statt:

x_i	$(3, 1, 7, 5, 3, 3)$
$\mathrm{rg}(x_i)$	$(a, 1, 6, 5, a, a)$

wobei der Durchschnittsrang a von Stichprobeneintrag 3 gleich $a = {}^1\!/_3(2+3+4) = 3$ ist.
Somit wird der sogenannte *Spearmans Korrelationskoeffizient* (Rangkorrelationskoeffizient) der Stichproben

$$(x_1, \ldots, x_n) \quad \text{und} \quad (y_1, \ldots, y_n)$$

als der *Bravais-Pearson-Koeffizient* der Stichproben ihrer Ränge

$$(\mathrm{rg}(x_1), \ldots, \mathrm{rg}(x_n)) \quad \text{und} \quad (\mathrm{rg}(y_1), \ldots, \mathrm{rg}(y_n))$$

definiert:

$$\varrho_{sp} = \frac{\sum_{i=1}^n \left(\mathrm{rg}(x_i) - \overline{\mathrm{rg}}_x\right)\left(\mathrm{rg}(y_i) - \overline{\mathrm{rg}}_y\right)}{\sqrt{\sum_{i=1}^n \left(\mathrm{rg}(x_i) - \overline{\mathrm{rg}}_x\right)^2 \sum_{i=1}^n \left(\mathrm{rg}(y_i) - \overline{\mathrm{rg}}_y\right)^2}},$$

wobei

$$\overline{\mathrm{rg}}_x = \frac{1}{n}\sum_{i=1}^n \mathrm{rg}(x_i) = \frac{1}{n}\sum_{i=1}^n \mathrm{rg}(x_{(i)}) = \frac{1}{n}\sum_{i=1}^n i = \frac{n(n+1)}{2n} = \frac{n+1}{2},$$
$$\overline{\mathrm{rg}}_y = \frac{1}{n}\sum_{i=1}^n \mathrm{rg}(y_i) = \frac{n+1}{2}.$$

Dieselbe Darstellung $\overline{\mathrm{rg}}_y$ gilt auch, wenn Bindungen vorhanden sind.

Dieser Koeffizient misst monotone Zusammenhänge in den Daten. Aus den Eigenschaften der Bravais-Pearson-Koeffizienten folgt $|\varrho_{sp}| \leq 1$. Betrachten wir die Fälle $\varrho_{sp} = \pm 1$ gesondert:

- $\varrho_{sp} = 1$ bedeutet, dass die Punkte $(\mathrm{rg}(x_i), \mathrm{rg}(y_i))$, $i = 1, \ldots, n$ auf einer Geraden mit positiver Steigung liegen. Da aber $\mathrm{rg}(x_i), \mathrm{rg}(y_i)$ Werte in $\mathbb{N}$ sind, kann diese Steigung nur 1 sein. Es bedeutet, dass dem kleinsten Wert in der Stichprobe $(x_1, \ldots, x_n)$ der kleinste Wert in $(y_1, \ldots, y_n)$ entspricht, usw., d.h., für wachsende x_i wachsen auch die y_i streng monoton: $x_i < x_j \Longrightarrow y_i < y_j \quad \forall i \neq j$.
- Analog gilt dann für $\varrho_{sp} = -1$, dass $x_i < x_j \Longrightarrow y_i > y_j \quad \forall i \neq j$.

Dies kann folgendermaßen zusammengefaßt werden:

- $\varrho_{sp} > 0$: gleichsinniger monotoner Zusammenhang (x_i groß $\Longleftrightarrow$ y_i groß)
- $\varrho_{sp} < 0$: gegensinniger monotoner Zusammenhang (x_i groß $\Longleftrightarrow$ y_i klein)
- $\varrho_{sp} \approx 0$: kein monotoner Zusammenhang.

Da der Spearmans Korrelationskoeffizient nur Ränge von x_i und y_i betrachtet, eignet er sich auch für ordinale (und nicht nur quantitative) Daten.

Lemma 7.9.2 Falls die Stichproben $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$ keine Bindung enthalten ($x_i \neq x_j$, $y_i \neq y_j$ $\forall i \neq j$), dann gilt

$$\varrho_{sp} = 1 - \frac{6}{(n^2-1)n} \sum_{i=1}^{n} d_i^2 \,,$$

wobei $d_i = \mathrm{rg}(x_i) - \mathrm{rg}(y_i) \quad \forall i = 1, \ldots, n.$

Beweis Als Übungsaufgabe. □

Satz 7.9.1 (Invarianzeigenschaften) 1. Wenn die Merkmale X und Y linear transformiert werden:

$$f(X) = a_x X + b_x \,, \quad a_x \neq 0,\, b_x \in \mathbb{R},$$
$$g(Y) = a_y Y + b_y \,, \quad a_y \neq 0,\, b_y \in \mathbb{R},$$

dann gilt $\varrho_{f(x)g(y)} = \mathrm{sgn}\ (a_x a_y) \cdot \varrho_{xy}$.

2. Falls Funktionen $f : \mathbb{R} \to \mathbb{R}$ und $g : \mathbb{R} \to \mathbb{R}$ beide monoton wachsend oder beide monoton fallend sind, dann gilt

$$\varrho_{sp}(f(x), g(y)) = \varrho_{sp}(x, y) \,.$$

Falls f monoton wachsend und g monoton fallend (oder umgekehrt) sind, dann gilt $\varrho_{sp}(f(x), g(y)) = -\varrho_{sp}(x, y)$.

Beweis Beweisen wir nur 1., weil 2. offensichtlich ist.

1. Es gilt

$$\begin{aligned}
&\varrho_{f(x)g(y)} \\
&= \frac{\sum_{i=1}^n \big((a_x x_i + b_x) - (a_x \bar{x}_n + b_x)\big)\big((a_y y_i + b_y) - (a_y \bar{y}_n + b_y)\big)}{\sqrt{a_x^2 \sum_{i=1}^n (x_i - \bar{x}_n)^2 a_y^2 \sum_{i=1}^n (y_i - \bar{y}_n)^2}} \\
&= \frac{a_x a_y}{|a_x||a_y|} \cdot \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sqrt{\sum_{i=1}^n (x_i - \bar{x}_n)^2 \sum_{i=1}^n (y_i - \bar{y}_n)^2}} = \mathrm{sgn}\ (a_x a_y) \cdot \varrho_{xy} \,.
\end{aligned}$$

□

Bemerkung 7.9.1 1. Da lineare Transformationen monoton sind, gilt Aussage 1) auch für Spearmans Korrelationskoeffizienten ϱ_{sp}.

2. Der Koeffizient ϱ_{xy} erfasst lineare Zusammenhänge, während ϱ_{sp} monotone Zusammenhänge aufspürt.

7.9.2 Einfache lineare Regression

Wenn man den Zusammenhang von Merkmalen X und Y mit Hilfe von Streudiagrammen visualisiert, wird oft ein linearer Trend erkennbar, obwohl der Bravais-Pearson-Korrelationskoeffizient einen Wert kleiner als 1 liefert, z.B. $\varrho_{xy} \approx 0,6$ (vgl. Abb. 7.9). Dies ist der Fall, weil die Datenpunkte (x_i, y_i),

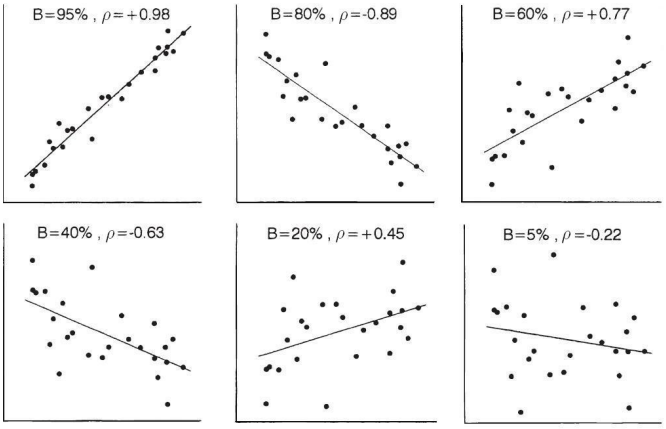


Abbildung 7.9: Vergleich verschiedenwertiger Bestimmtheitsmaße. Es sind Regressionsgerade, Bestimmtheitsmaß B und Korrelationskoeffizient ρ verschiedener (fiktiver) Punktwolken vom Umfang $n = 25$ dargestellt. Die Beschriftung der Achsen ist weggelassen, weil sie hier ohne Bedeutung ist.

$i = 1, \ldots, n$ oft um eine Gerade streuen und nicht exakt auf einer Geraden liegen. Um solche Situationen stochastisch modellieren zu können, nimmt man den Zusammenhang der Form

$$Y = f(X) + \varepsilon$$

an, wobei ε die sogenannte Störgröße ist, die auf mehrere Ursachen wie z.B. Beobachtungsfehler (Messfehler, Berechnungsfehler, usw.) zurückzuführen sein kann. Dabei nennt man die Zufallsvariable Y *Zielgröße* oder *Regressand*, die Zufallsvariable X *Einflussfaktor*, *Regressor* oder *Ausgangsvariable*. Der Zusammenhang $Y = f(X) + \varepsilon$ wird *Regression* genannt, wobei man oft über ε voraussetzt, dass $\mathrm{E}\varepsilon = 0$ (kein systematischer Beobachtungsfehler). Wenn $f(x) = \alpha + \beta x$ eine lineare Funktion ist, so spricht man von der *einfachen linearen Regression*. Es sind aber durchaus andere Arten der Zusammenhänge denkbar, wie z.B.

$$f(x) = \sum_{i=0}^{n} \alpha_i x^i$$

X	Y
Geschwindigkeit	Länge des Bremswegs
Körpergröße des Vaters	Körpergröße des Sohnes
Produktionsfaktor	Qualität des Produktes
Spraydosen-Verbrauch	Ozongehalt der Atmosphäre
Noten im Bachelor-Studium	Noten im Master-Studium

Tabelle 7.1: Beispiele möglicher Ausgangs- und Zielgrößen

(*polynomiale Regression*), usw. Beispiele für mögliche Ausgangs- bzw. Zielgrößen sind in Tabelle 7.1 zusammengefasst, einige Beispiele in Abbildung 7.10.

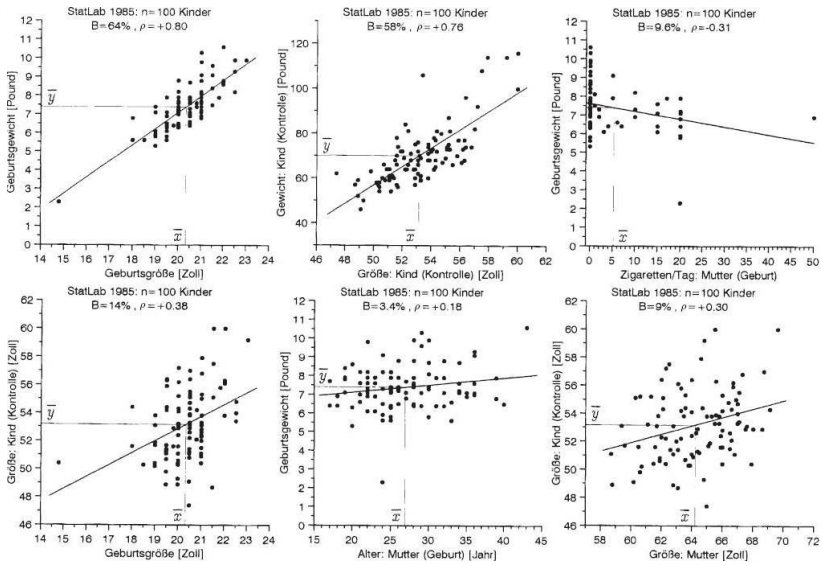


Abbildung 7.10: Punktwolken verschiedener Merkmale der StatLab-Auswahl 1985 mit Regressionsgerade, Bestimmtheitsmaß B und Korrelationskoeffizient ϱ.

Auf Modellebene ist damit folgende Fragestellung gegeben: Gegeben seien Zufallsstichproben von Ziel- bzw. Ausgangsvariablen $(Y_1, \ldots, Y_n)$ und $(X_1, \ldots, X_n)$, zwischen denen ein verrauschter linearer Zusammenhang $Y_i = \alpha + \beta X_i + \varepsilon_i$ besteht, wobei ε_i Störgrößen sind, die nicht direkt beobachtbar und uns somit unbekannt sind. Meistens nimmt man an, dass $\mathrm{E}\,\varepsilon_i = 0 \quad \forall\, i = 1, \ldots, n$ und $\mathrm{Cov}\,(\varepsilon_i, \varepsilon_j) = \sigma^2 \delta_{ij}$, d.h. $\varepsilon_1 \ldots \varepsilon_n$ sind unkorreliert mit $\mathrm{Var}\ \varepsilon_i = \sigma^2$. Wenn wir über die Eigenschaften der Schätzer für α, β

und σ^2 reden, gehen wir davon aus, dass die X-Werte nicht zufällig sind, also $X_i = x_i \quad \forall i = 1, \ldots, n$. Wenn man von einer konkreten Stichprobe $(y_1, \ldots, y_n)$ für $(Y_1, \ldots, Y_n)$ ausgeht, so sollen anhand von den Stichproben $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$ Regressionsparameter α (*Regressionskonstante*) und β (*Regressionskoeffizient*) sowie *Regressionsvarianz* σ^2 geschätzt werden. Dabei verwendet man die sogenannte *Methode der kleinsten Quadrate*, die den mittleren quadratischen Fehler von den Datenpunkten $(x_i, y_i)_{i=1,\ldots,n}$ des Streudiagramms zur *Regressionsgeraden* $y = \alpha + \beta x$ minimiert:

$$(\alpha, \beta) = \operatorname*{arg\,min}_{\alpha,\beta \in \mathbb{R}} e(\alpha, \beta) \quad \text{mit} \quad e(\alpha, \beta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \,.$$

Da die Darstellung $y_i = \alpha + \beta x_i + \varepsilon_i$ gilt, kann man $e(\alpha, \beta) = 1/n \sum_{i=1}^{n} \varepsilon_i^2$

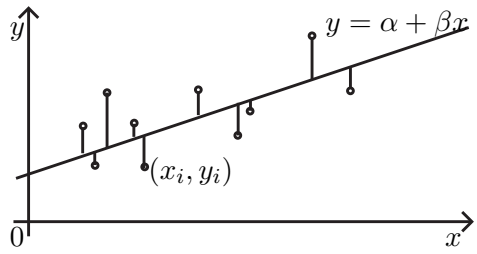


Abbildung 7.11: Methode kleinster Quadrate

schreiben. Es ist der vertikale mittlere quadratische Abstand von den Datenpunkten (x_i, y_i) zur Geraden $y = \alpha + \beta x$ (vgl. Abb. 7.11). Das Minimierungsproblem $e(\alpha, \beta) \mapsto \min$ löst man durch das zweifache Differenzieren von $e(\alpha, \beta)$. Somit erhält man $\hat{\alpha} = \bar{y}_n - \hat{\beta}\bar{x}_n$, wobei

$$\hat{\beta} = \frac{S_{xy}^2}{S_{xx}^2} \,, \quad \bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i \,, \quad \bar{y}_n = \frac{1}{n} \sum_{i=1}^{n} y_i \,,$$

$$S_{xy}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)(y_i - \bar{y}_n) \,, \quad S_{xx}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \,.$$

Übungsaufgabe 7.9.1 Leiten Sie die Schätzer $\hat{\alpha}$ und $\hat{\beta}$ selbstständig her.

Die Varianz σ^2 schätzt man durch $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2$, wobei $\hat{\varepsilon}_i = y_i - \hat{\alpha} - \hat{\beta} x_i \,, \quad i = 1, \ldots, n$ die sogenannten *Residuen* sind. Die Gründe, warum $\hat{\sigma}^2$ diese Gestalt hat, können an dieser Stelle noch nicht angegeben werden, weil wir noch nicht die Maximum-Likelihood-Methode kennen. Zu gegebener Zeit (in der Vorlesung Stochastik III) wird jedoch klar, dass diese Art der Schätzung sehr natürlich ist.

Bemerkung 7.9.2 Die angegebenen Schätzer für α und β sind nicht symmetrisch bzgl. Variablen x_i und y_i. Wenn man also die *horizontalen* Abstände (statt vertikaler) zur Bildung des mittleren quadratischen Fehlers nimmt

Kind i	1	2	3	4	5	6	7	8	9
Fernsehzeit x_i	0,3	2,2	0,5	0,7	1,0	1,8	3,0	0,2	2,3
Tiefschlafdauer y_i	5,8	4,4	6,5	5,8	5,6	5,0	4,8	6,0	6,1

Tabelle 7.2: Daten von Fernsehzeit und korrespondierender Tiefschlafdauer

(was dem Rollentausch $x \leftrightarrow y$ entspricht), so bekommt man andere Schätzer für α und β, die mit $\hat{\alpha}$ und $\hat{\beta}$ nicht übereinstimmen müssen:

$$d_i = y_i - \alpha - \beta x_i \mapsto d_i' = x_i - \frac{(y_i - \alpha)}{\beta}\,.$$

Ein Ausweg aus dieser asymmetrischen Situation wäre es, die orthogonalen

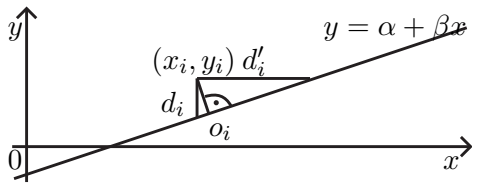


Abbildung 7.12: Orthogonale Abstände

Abstände o_i von (x_i, y_i) zur Geraden $y = \alpha + \beta x$ zu betrachten (vgl. Abb. 7.12). Diese Art der Regression, die „errors-in-variables regression“ genannt wird, hat aber eine Reihe von Eigenschaften, die sie zur Prognose von Zielvariablen y_i durch die Ausgangsvariablen x_i unbrauchbar machen. Sie sollte zum Beispiel nur dann verwendet werden, wenn die Standardabweichungen für X und Y etwa gleich groß sind.

Beispiel 7.9.1 Ein Kinderpsychologe vermutet, dass sich häufiges Fernsehen negativ auf das Schlafverhalten von Kindern auswirkt. Um diese Hypothese zu überprüfen, wurden 9 Kinder im gleichen Alter befragt, wie lange sie pro Tag fernsehen dürfen, und zusätzlich die Dauer ihrer Tiefschlafphase gemessen. So ergibt sich der Datensatz in Tabelle 7.2 und die Regressionsgerade aus Abbildung 7.13.

Es ergibt sich für die oben genannten Stichproben $(x_1, \ldots, x_9)$, $(y_1, \ldots, y_9)$

$$\bar{x}_9 = 1,33\,, \quad \bar{y}_9 = 5,56\,, \quad \hat{\beta} = -0,45\,, \quad \hat{\alpha} = 6,16\,.$$

Somit ist

$$y = 6,16 - 0,45x$$

die Regressionsgerade, die eine negative Steigung hat, was die Vermutung des Kinderpsychologen bestätigt. Außerdem ist es mit Hilfe dieser Geraden

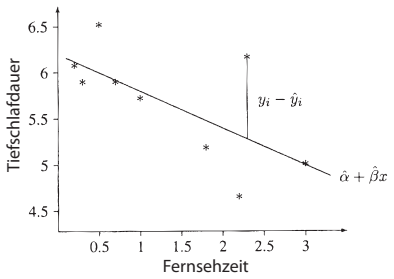


Abbildung 7.13: Streudiagramm und Ausgleichsgerade zur Regression der Dauer des Tiefschlafs auf die Fernsehzeit

möglich, Prognosen für die Dauer des Tiefschlafs für vorgegebene Fernsehzeiten anzugeben. So wäre z.B. für die Fernsehzeit von 1 Stunde der Tiefschlaf von $6,16 - 0,45 \cdot 1 = 5,71$ Stunden plausibel.

Bemerkung 7.9.3 (Eigenschaften der Regressionsgerade) 1. Für $\hat{\beta}$ gilt, dass sgn $(\hat{\beta})$ = sgn (ρ_{xy}), was aus $\hat{\beta} = s^2_{xy}/s^2_{xx}$ folgt. Dies bedeutet (falls $s^2_{yy} > 0$):

   (a) Die Regressionsgerade $y = \hat{\alpha} + \hat{\beta}x$ steigt an, falls die Stichproben $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$ positiv korreliert sind.

   (b) Die Regressionsgerade fällt ab, falls sie negativ korreliert sind.

   (c) Die Regressionsgerade ist konstant, falls die Stichproben unkorreliert sind.

   Falls $s^2_{yy} = 0$, dann ist die Regressionsgerade konstant $(y = \bar{y}_n)$.

2. Die Regressionsgerade $y = \hat{\alpha} + \hat{\beta}x$ verläuft immer durch den Punkt $(\bar{x}_n, \bar{y}_n)$: $\hat{\alpha} + \hat{\beta}\bar{x}_n = \bar{y}_n$.

3. Seien $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$, $i = 1, \ldots, n$. Dann gilt

$$\overline{\hat{y}_n} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = \bar{y}_n \quad \text{und somit} \quad \sum_{i=1}^{n} \underbrace{(y_i - \hat{y}_i)}_{\hat{\varepsilon}_i} = 0\,.$$

   Dabei sind $\hat{\varepsilon}_i$ die schon vorher eingeführten Residuen. Mit ihrer Hilfe ist es möglich, die Güte der Regressionsprognose zu beurteilen.

Residualanalyse und Bestimmtheitsmaß

Definition 7.9.1 Der relative Anteil der Streuungsreduktion an der Gesamtstreuung S^2_{yy} heißt das *Bestimmtheitsmaß* der Regressionsgeraden:

$$R^2 = \frac{S^2_{yy} - \frac{1}{n-1}\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{S^2_{yy}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_n)^2}\,.$$

Es ist nur im Fall $S_{xx}^2 > 0$, $S_{yy}^2 > 0$ definiert, d.h., wenn nicht alle Werte x_i bzw. y_i übereinstimmen.

Warum R^2 in dieser Form eingeführt wird, zeigt folgende Überlegung, die *Streuungszerlegung* genannt wird:

Lemma 7.9.3 Die Gesamtstreuung („sum of squares total“)

$$\mathrm{SQT} = (n-1)S_{yy}^2 = \sum_{i=1}^{n}(y_i - \bar{y}_n)^2$$

lässt sich in die Summe der sogenannten erklärten Streuung „sum of squares explained“ $\mathrm{SQE} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_n)^2$ und der Residualstreuung „sum of squared residuals“ $\mathrm{SQR} = \sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ zerlegen:

$$\mathrm{SQT} = \mathrm{SQE} + \mathrm{SQR}$$

bzw.

$$\sum_{i=1}^{n}(y_i - \bar{y}_n)^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_n)^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \,.$$

Die erklärte Streuung gibt die Streuung der Regressionsgeradenwerte um $\bar{y}_n$ an. Sie stellt damit die auf den linearen Zusammenhang zwischen X und Y zurückgeführende Variation der y-Werte dar. Das oben eingeführte Bestimmtheitsmaß ist somit der Anteil dieser Streuung an der Gesamtstreuung:

$$R^2 = \frac{\mathrm{SQE}}{\mathrm{SQT}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_n)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_n)^2} = \frac{\mathrm{SQT} - \mathrm{SQR}}{\mathrm{SQT}} = 1 - \frac{\mathrm{SQR}}{\mathrm{SQT}} \,.$$

Es folgt aus dieser Darstellung, dass $R^2 \in [0,1]$ ist.

1. $R^2 = 0$ bedeutet $\mathrm{SQE} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_n)^2 = 0$ und somit $\hat{y}_i = \bar{y}_n \ \forall i$. Dies weist darauf hin, dass das lineare Modell in diesem Fall schlecht ist, denn aus $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i = \bar{y}_n$ folgt $\hat{\beta} = \frac{S_{xy}^2}{S_{xx}^2} = 0$ und somit $S_{xy}^2 = 0$. Also sind die Merkmale X und Y unkorreliert.

2. $R^2 = 1$ bedingt $\mathrm{SQR} = \sum_{i=1}^{n}\hat{\varepsilon}_i^2 = 0$. Somit liegen alle (x_i, y_i) perfekt auf der Regressionsgeraden. Dies bedeutet, dass die Daten x_i und y_i, $i = 1, \ldots, n$ perfekt linear abhängig sind.

   Faustregel zur Beurteilung der Güte der Anpassung eines linearen Modells an Hand von Bestimmtheitsmaß R^2:
   R^2 ist deutlich von Null verschieden (d.h. es besteht noch ein linearer Zusammenhang), falls $R^2 > \frac{4}{n+2}$, wobei n der Stichprobenumfang ist.

Allgemein gilt folgender Zusammenhang zwischen dem Bestimmtheitsmaß R^2 und dem Bravais-Pearson-Korrelationskoeffizienten ϱ_{xy}:

Lemma 7.9.4

$$R^2 = \varrho_{xy}^2$$

Folgerung 7.9.1 1. Der Wert von R^2 ändert sich bei einer Lineartransformation der Daten $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$ nicht. Grafisch kann man die Güte der Modellanpassung bei der linearen Regression folgendermaßen überprüfen:

Man zeichnet Punktepaare $(\hat{y}_i, \hat{\varepsilon}_i)_{i=1,\ldots,n}$ als Streudiagramm (der sogenannte *Residualplot*). Falls diese Punktewolke gleichmäßig um Null streut, so ist das lineare Modell gut gewählt worden. Falls das Streudiagramm einen erkennbaren Trend aufweist, bedeutet das, dass die Annahme des linearen Modells für diese Daten ungeeigenet sei (vgl. Abb. 7.14)

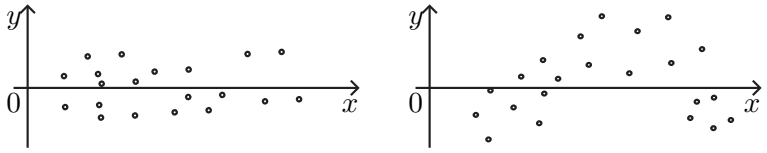


Abbildung 7.14: Links: Gute, Rechts: Schlechte Übereinstimmung mit dem linearen Modell

2. Da $R^2 = \varrho_{xy}^2$, ist der Wert von R^2 symmetrisch bzgl. der Stichproben $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$:

$$\varrho_{xy}^2 = R^2 = \varrho_{yx}^2 \qquad \text{bzw.} \qquad R_{xy}^2 = R_{yx}^2\,,$$

wobei R_{xy}^2 das Bestimmtheitsmaß bezeichnet, das sich aus der normalen Regression ergibt und R_{yx}^2 das mit vertauschten Achsen.

Kapitel 8

Punktschätzer

8.1 Parametrisches Modell

Sei $(x_1, \ldots, x_n)$ eine konkrete Stichprobe. Es wird angenommen, dass die Stichprobe $(x_1, \ldots, x_n)$ eine Realisierung einer Zufallsstichprobe $(X_1, \ldots, X_n)$ ist, wobei $X_1, \ldots, X_n$ unabhängige identisch verteilte Zufallsvariablen mit der unbekannten Verteilungsfunktion F sind und F zu einer bekannten parametrischen Familie $\{F_\theta : \theta \in \Theta\}$ gehört. Hier ist $\theta = (\theta_1, \ldots, \theta_m) \in \Theta$ der *m-dimensionale Parametervektor* der Verteilung F_θ und $\Theta \subset \mathbb{R}^m$ der sogenannte *Parameterraum* (eine Borel-Teilmenge von $\mathbb{R}^m$, die die Menge aller zugelassenen Parameterwerte darstellt). Es wird vorausgesetzt, dass die Parametrisierung $\theta \to F_\theta$ *identifizierbar* ist, indem $F_{\theta_1} \neq F_{\theta_2}$ für $\theta_1 \neq \theta_2$ gilt.

Eine wichtige Aufgabe der Statistik, die wir in diesem Kapitel betrachten werden, besteht in der Schätzung des Parametervektors θ (oder eines Teils von θ) an Hand von der konkreten Stichprobe $(x_1, \ldots, x_n)$. In diesem Fall spricht man von einem *Punktschätzer* $\hat{\theta} : \mathbb{R}^n \to \mathbb{R}^m$, der eine gültige Stichprobenfunktion ist. Meistens wird angenommen, dass

$$P\left(\hat{\theta}(X_1, \ldots, X_n) \in \Theta\right) = 1\,,$$

wobei es zu dieser Regel auch Ausnahmen gibt.

Beispiel 8.1.1 1. Sei X die Dauer des fehlerfreien Arbeitszyklus eines technischen Systems. Oft wird $X \sim Exp(\lambda)$ angenommen. Dann stellt $\{F_\theta : \theta \in \Theta\}$ mit $m = 1$, $\theta = \lambda$, $\Theta = \mathbb{R}_+$ und

$$F_\theta(x) = (1 - e^{-\theta x}) \cdot I(x \geq 0)$$

ein parametrisches Modell dar, wobei der Parameterraum eindimensional ist. Später wird für λ der (Punkt-) Schätzer $\hat{\lambda}(x_1, \ldots, x_n) = 1/\bar{x}_n$ vorgeschlagen.

2. In den Fragestellungen der statistischen Qualitätskontrolle werden n Erzeugnisse auf Mängel untersucht. Falls $p \in (0,1)$ die unbekannte Wahrscheinlichkeit des Mangels ist, so wird mit $X \sim Bin(n,p)$ die Gesamtanzahl der mangelhaften Produkte beschrieben. Dabei wird folgendes parametrische Modell unterstellt:

$$\Theta = \{(n,p) : n \in \mathbb{N},\, p \in (0,1)\}\,, \qquad \theta = (n,p)\,, \quad m = 2\,,$$

$$F_\theta(x) = P_\theta(X \le x) = \begin{cases} 1, & x > n \\ \sum_{k=0}^{[x]} \binom{n}{k} p^k (1-p)^{n-k}, & x \in [0,n] \\ 0, & x < 0. \end{cases}$$

Falls n bekannt ist, kann die Wahrscheinlichkeit p des Ausschusses durch den Punktschätzer $\hat{p}(x_1,\ldots,x_n) = \bar{x}_n\,, \quad x_i \in \{0,1\}$ näherungsweise berechnet werden.

8.2 Parametrische Familien von statistischen Prüfverteilungen

In der Vorlesung Wahrscheinlichkeitsrechung wurden bereits einige parametrische Familien von Verteilungen eingeführt. Hier geben wir weitere Verteilungsfamilien an, die in der Statistik eine besondere Stellung einnehmen, weil sie als Referenzverteilungen in der Schätztheorie, statistischen Tests und Vertrauensintervallen ihre Anwendung finden.

8.2.1 Gamma-Verteilung

Als erstes führen wir zwei spezielle Funktionen aus der Analysis ein:

1. Die *Gamma-Funktion:*

$$\Gamma(p) = \int_0^\infty x^{p-1} e^{-x} dx \qquad \text{für } p > 0\,.$$

Es gelten folgende Eigenschaften:

$$\Gamma(1) = 1\,, \qquad \Gamma(1/2) = \sqrt{\pi}$$
$$\Gamma(p+1) = p\Gamma(p) \qquad \forall\, p > 0\,, \qquad \Gamma(n+1) = n!\,, \qquad \forall\, n \in \mathbb{N}\,.$$

2. Die *Beta-Funktion:*

$$B(p,q) = \int_0^1 t^{p-1}(1-t)^{q-1}\, dt\,, \quad p,q > 0\,.$$

Es gelten folgende Eigenschaften:

$$B(p,q) = B(q,p)\,, \qquad B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}\,, \qquad p,q > 0\,.$$

Definition 8.2.1 Die *Gamma-Verteilung* mit Parametern $\lambda > 0$ und $p > 0$ ist eine absolut stetige Verteilung mit der Dichte

$$f_X(x) = \begin{cases} \frac{\lambda^p x^{p-1}}{\Gamma(p)} e^{-\lambda x}\,, & x \geq 0\,, \\ 0\,, & x < 0\,. \end{cases} \tag{8.1}$$

Dabei verwenden wir die Bezeichnung $X \sim \Gamma(\lambda, p)$ für eine Zufallsvariable X, die Gamma-verteilt mit Parametern λ und p ist. Es gilt offensichtlich $X \geq 0$ fast sicher für $X \sim \Gamma(\lambda, p)$.

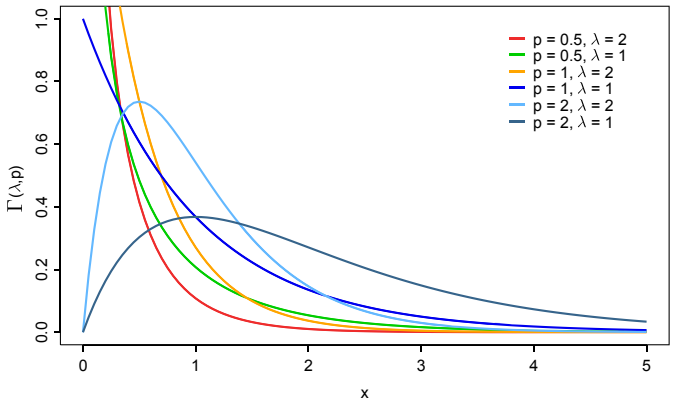


Abbildung 8.1: Dichte der Gammaverteilung

Übungsaufgabe 8.2.1 Zeigen Sie, dass (8.1) eine Dichte ist.

Beispiel 8.2.1

1. In der Kraftfahrzeugversicherung wird die Gamma-Verteilung oft zur Modellierung des Gesamtschadens verwendet.
2. Falls $p = 1$, dann ist $\Gamma(\lambda, 1) = Exp(\lambda)$.

Satz 8.2.1 (Momenterzeugende und charakteristische Funktion der Gammaverteilung) Falls $X \sim \Gamma(\lambda, p)$, dann gilt Folgendes:

1. Die momenterzeugende Funktion der Gammaverteilung $\Psi_X(s)$ ist gegeben durch

$$\Psi_X(s) = \mathrm{E}e^{sX} = \frac{1}{(1 - s/\lambda)^p}\,, \quad s < \lambda\,.$$

2. k-te Momente:

$$\mathrm{E}X^k = \frac{p(p+1)\cdot\ldots\cdot(p+k-1)}{\lambda^k}\,, \quad k \in \mathbb{N}\,.$$

Folgerung 8.2.1 (Faltungsstabilität der Γ-Verteilung) Falls $X \sim \Gamma(\lambda, p_1)$ und $Y \sim \Gamma(\lambda, p_2)$, X, Y unabhängig, dann ist $X + Y \sim \Gamma(\lambda, p_1 + p_2)$.

Beweis Es gilt

$$\begin{aligned}\varphi_{X+Y}(s) &= \varphi_X(s) \cdot \varphi_Y(s) \\ &= \frac{1}{(1-is/\lambda)^{p_1}} \cdot \frac{1}{(1-is/\lambda)^{p_2}} = \left(\frac{1}{1-is/\lambda}\right)^{p_1+p_2} \\ &= \varphi_{\Gamma(\lambda,p_1+p_2)}(s)\,.\end{aligned}$$

Da die charakteristischen Funktionen die Verteilungen eindeutig bestimmen, folgt damit $X + Y \sim \Gamma(\lambda, p_1 + p_2)$. □

Beispiel 8.2.2 Seien $X_1, \ldots, X_n \sim Exp(\lambda)$ unabhängig. Nach der Folgerung 8.2.1 gilt $X = X_1 + \ldots + X_n \sim \Gamma(\lambda, \underbrace{1 + \ldots + 1}_{n}) = \Gamma(\lambda, n)$, denn $Exp(\lambda) = \Gamma(\lambda, 1)$. Dabei heißt X *Erlang-verteilt* mit Parametern λ und n. Man schreibt $X \sim Erl(\lambda, n)$.

Zusammengefasst: $Erl(\lambda, n) = \Gamma(\lambda, n)$

Interpretation: In der Risikotheorie z.B. sind X_i Zwischenankunftszeiten der Einzelschäden. Dann ist $X = \sum_{i=1}^{n} X_i$ die Ankunftszeit des n-ten Schadens, $X \sim Erl(\lambda, n)$.

Definition 8.2.2 (χ^2-Verteilung) X ist eine χ^2-verteilte Zufallsvariable mit k Freiheitsgraden ($X \sim \chi^2_k$), falls $X \stackrel{d}{=} X_1^2 + \ldots + X_k^2$, wobei $X_1, \ldots, X_k \sim N(0,1)$ unabhängige identisch verteilte Zufallsvariablen sind.

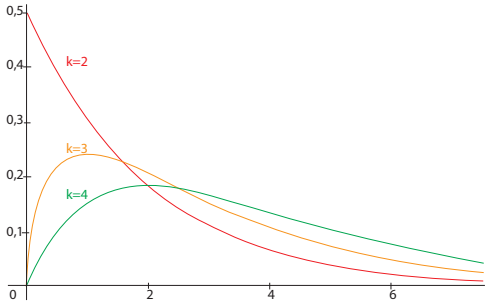


Abbildung 8.2: Dichte der χ^2-Verteilung für $k = 2, 3, 4$

Satz 8.2.2 (χ^2-Verteilung: Spezialfall der Γ-Verteilung mit $\lambda = 1/2$, $p = k/2$) Falls $X \sim \chi^2_k$, dann gilt:

1. $X \sim \Gamma(1/2, k/2)$, d.h.

$$f_X(x) = \begin{cases} \dfrac{x^{k/2-1}e^{-x/2}}{2^{k/2}\Gamma(k/2)}\,, & x \geq 0 \\ 0, & x < 0 \end{cases}. \tag{8.2}$$

2. Insbesondere ist $\mathrm{E}X = k$, $\mathrm{Var}\ X = 2k$.

8.2.2 Student-Verteilung (t-Verteilung)

Definition 8.2.3 Seien X, Y unabhängige Zufallsvariablen, wobei $X \sim N(0,1)$ und $Y \sim \chi^2_r$. Dann heißt die Zufallsvariable

$$U \stackrel{d}{=} \frac{X}{\sqrt{Y/r}}$$

Student- oder *t-verteilt mit r Freiheitsgraden*. Wir schreiben $U \sim t_r$.

Satz 8.2.3 (Dichte der t-Verteilung) Falls $X \sim t_r$, dann gilt:

1. $$f_X(x) = \frac{1}{\sqrt{r}B\left(\frac{r}{2}, \frac{1}{2}\right)} \cdot \frac{1}{\left(1 + \frac{x^2}{r}\right)^{\frac{r+1}{2}}}, \qquad x \in \mathbb{R}.$$

2. $\mathrm{E}X = 0$, $\quad \mathrm{Var}\ X = \frac{r}{r-2}$, $\quad r \geq 3$.

Bemerkung 8.2.1 1. **Grafik von f_r:** Die t_r-Verteilung ist symmetrisch.

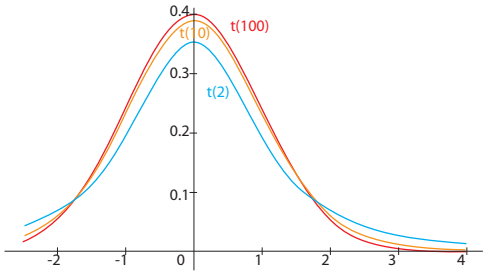


Abbildung 8.3: Dichte f_r der t-Verteilung für $r = 2, 10, 100$

Insbesondere gilt:

$$t_{r,\alpha} = -t_{r,1-\alpha}, \quad \alpha \in (0,1),$$

wobei $t_{r,\alpha}$ das α-Quantil der Student-Verteilung mit r Freiheitsgraden ist.

2. Falls $r \to \infty$, dann $f_r(x) \to \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, $\quad x \in \mathbb{R}$. (Übungsaufgabe)

3. Für $r = 1$ gilt: $t_1 = Cauchy(0,1)$ mit Dichte $f(x) = \frac{1}{\pi(1+x^2)}$. Der Erwartungswert von t_1 existiert nicht.

8.2.3 Fisher-Snedecor-Verteilung (F-Verteilung)

Definition 8.2.4 Falls $X \stackrel{d}{=} \frac{U_r/r}{U_s/s}$, wobei $U_r \sim \chi^2_r$, $\quad U_s \sim \chi^2_s$, $r, s \in \mathbb{N}$, U_r, U_s unabhängig, dann hat X eine F-Verteilung mit Freiheitsgraden r, s. Bezeichnung: $X \sim F_{r,s}$.

Bemerkung 8.2.2 Sei $X \sim F_{r,s},\ r, s \in \mathbb{N}$ mit Dichte f_X.

1. Einige Graphen der Dichte der F-Verteilung sind in Abbildung 8.4 dargestellt.

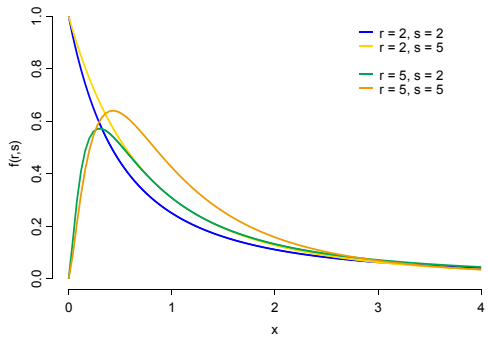


Abbildung 8.4: Dichte F_X der F-Verteilung für $r = 2, 5$ und $s = 2, 5$.

2. Einige Eigenschaften der F-Verteilung:

   Lemma 8.2.1 Es gilt:

   (a)
   $$\mathrm{E}X = \frac{s}{s-2}, \qquad s \geq 3\,.$$

   (b)
   $$\mathrm{Var}\ X = \frac{2s^2(r+s-2)}{r(s-4)(s-2)^2}\,, \qquad s \geq 5\,.$$

   (c) Falls $F_{r,s,\alpha}$ das α-Quantil der $F_{r,s}$-Verteilung ist, dann gilt
   $$F_{r,s,\alpha} = \frac{1}{F_{s,r,1-\alpha}}\,, \quad \alpha \in (0,1)\,.$$

8.3 Punktschätzer und ihre Grundeigenschaften

Sei $(X_1, \ldots, X_n)$ eine Zufallsstichprobe, definiert auf dem kanonischen Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, P_\theta)$. Seien $X_i,\ i = 1, \ldots, n$ unabhängige identisch verteilte Zufallsvariablen mit Verteilungsfunktion $F \in \{F_\theta : \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^m$. Finde einen Schätzer $\hat{\theta}(X_1, \ldots, X_n)$ für den Parameter θ mit vorgegebenen Eigenschaften.

Unser Ziel im nächsten Abschnitt ist es, zunächst grundlegende Eigenschaften der Schätzer kennenzulernen.

8.3.1 Eigenschaften von Punktschätzern

Definition 8.3.1 (Erwartungstreue) Ein Schätzer $\hat{\theta}(X_1, \ldots, X_n)$ für θ heißt *erwartungstreu* oder *unverzerrt*, falls

$$\mathbb{E}_\theta \hat{\theta}(X_1, \ldots, X_n) = \theta\,, \qquad \theta \in \Theta\,.$$

Dabei wird vorausgesetzt, dass

$$\mathrm{E}_\theta |\hat{\theta}(X_1, \ldots, X_n)| < \infty\,, \qquad \theta \in \Theta\,.$$

Der *Bias* (*Verzerrung*) eines Schätzers $\hat{\theta}(X_1, \ldots, X_n)$ ist gegeben durch

$$\mathrm{Bias}(\hat{\theta}) = \mathrm{E}_\theta\, \hat{\theta}(X_1, \ldots, X_n) - \theta\,.$$

Falls $\hat{\theta}(X_1, \ldots, X_n)$ erwartungstreu ist, dann gilt $\mathrm{Bias}(\hat{\theta}) = 0$ (kein systematischer Schätzfehler).

Definition 8.3.2 (Asymptotische Erwartungstreue) Es gilt, dass der Schätzer $\hat{\theta}(X_1, \ldots, X_n)$ für θ *asymptotisch erwartungstreu* (oder *asymptotisch unverzerrt*) ist, falls (für große Datenmengen)

$$\mathrm{E}_\theta\, \hat{\theta}(X_1, \ldots, X_n) \underset{n\to\infty}{\longrightarrow} \theta, \qquad \theta \in \Theta.$$

Definition 8.3.3 (Konsistenz) Falls

$$\hat{\theta}(X_1, \ldots, X_n) \underset{n\to\infty}{\longrightarrow} \theta, \qquad \theta \in \Theta$$

in L^2, stochastisch bzw. fast sicher, dann heißt der Schätzer $\hat{\theta}(X_1, \ldots, X_n)$ ein *konsistenter Schätzer* für θ im *mittleren quadratischen, schwachen bzw. starken Sinne.*

- $\hat{\theta}$ *L^2-konsistent*: für $\mathbb{E}_\theta\, \hat{\theta}^2(X_1, \ldots, X_n) < \infty$ gilt

$$\hat{\theta} \underset{n\to\infty}{\overset{L^2}{\longrightarrow}} \theta \Longleftrightarrow \mathrm{E}_\theta |\hat{\theta}(X_1, \ldots, X_n) - \theta|^2 \underset{n\to\infty}{\longrightarrow} 0\,, \qquad \theta \in \Theta\,.$$

- $\hat{\theta}$ *schwach konsistent*:

$$\hat{\theta} \underset{n\to\infty}{\overset{P}{\longrightarrow}} \theta \Longleftrightarrow P_\theta(|\hat{\theta}(X_1, \ldots, X_n) - \theta| > \varepsilon) \underset{n\to\infty}{\longrightarrow} 0\,, \qquad \varepsilon > 0\,, \quad \theta \in \Theta.$$

- $\hat{\theta}$ *stark konsistent*:

$$\hat{\theta} \underset{n\to\infty}{\overset{\text{f.s.}}{\longrightarrow}} \theta \Longleftrightarrow P_\theta\left(\lim_{n\to\infty} \hat{\theta}(X_1, \ldots, X_n) = \theta\right) = 1\,, \qquad \theta \in \Theta.$$

Daraus ergibt sich folgendes Diagramm (vgl. Wahrscheinlichkeitsrechungsskript, Kapitel 6).

$$\boxed{L^2 - \text{Konsistenz}} \Longrightarrow \boxed{\text{schwache Konsistenz}} \Longleftarrow \boxed{\text{starke Konsistenz}}$$

Definition 8.3.4 (Mittlerer quadratischer Fehler (mean squared error)) Der mittlere quadratische Fehler eines Schätzers $\hat{\theta}(X_1, \ldots, X_n)$ für θ ist definiert als

$$MSE(\hat{\theta}) = \mathrm{E}_\theta \left|\hat{\theta}(X_1, \ldots, X_n) - \theta\right|^2 .$$

Lemma 8.3.1 Falls $m = 1$ und $\mathrm{E}_\theta\, \hat{\theta}^2(X_1, \ldots, X_n) < \infty\,, \quad \theta \in \Theta$, dann gilt

$$MSE(\hat{\theta}) = \mathrm{Var}_{\,\theta}\, \hat{\theta} + \left(\mathrm{Bias}(\hat{\theta})\right)^2 .$$

Beweis

$$\begin{aligned} MSE(\hat{\theta}) &= \mathrm{E}_\theta(\hat{\theta} - \theta)^2 = \mathrm{E}_\theta(\hat{\theta} - \mathrm{E}_\theta\hat{\theta} + \mathrm{E}_\theta\hat{\theta} - \theta)^2 \\ &= \underbrace{\mathrm{E}_\theta(\hat{\theta} - \mathrm{E}_\theta\, \hat{\theta})^2}_{\mathrm{Var}_{\,\theta}\, \hat{\theta}} + 2\underbrace{\mathrm{E}_\theta(\hat{\theta} - \mathrm{E}_\theta\, \hat{\theta})}_{=0}\underbrace{(\mathrm{E}_\theta\, \hat{\theta} - \theta)}_{=const} + \underbrace{(\mathrm{E}_\theta\, \hat{\theta} - \theta)^2}_{=\mathrm{Bias}(\hat{\theta})^2} \\ &= \mathrm{Var}_{\,\theta}\, \hat{\theta} + \Big(\mathrm{Bias}(\hat{\theta})\Big)^2 . \end{aligned}$$

□

Bemerkung 8.3.1 Falls $\hat{\theta}$ *erwartungstreu* für θ ist, dann gilt $MSE(\hat{\theta}) = \mathrm{Var}_{\,\theta}\, \hat{\theta}$.

Definition 8.3.5 (Vergleich von Schätzern) Es seien $\hat{\theta}_1(X_1, \ldots, X_n)$ und $\hat{\theta}_2(X_1, \ldots, X_n)$ zwei Schätzer für θ. Man sagt, dass $\hat{\theta}_1$ *besser* ist als $\hat{\theta}_2$, falls

$$MSE(\hat{\theta}_1) < MSE(\hat{\theta}_2)\,, \qquad \theta \in \Theta\,.$$

Falls $m = 1$ und die Schätzer $\hat{\theta}_1$, $\hat{\theta}_2$ erwartungstreu sind, so ist $\hat{\theta}_1$ *besser als* $\hat{\theta}_2$, falls $\hat{\theta}_1$ die kleinere Varianz besitzt. Dabei wird stets vorausgesetzt, dass $\mathrm{E}_\theta\, \hat{\theta}_i^2 < \infty\,, \quad \theta \in \Theta$.

Definition 8.3.6 (Asymptotische Normalverteiltheit) Sei $\hat{\theta}(X_1, \ldots, X_n)$ ein Schätzer für θ $(m = 1)$. Falls $0 < \mathrm{Var}_{\,\theta}\, \hat{\theta}(X_1, \ldots, X_n) < \infty\,, \quad \theta \in \Theta$ und

$$\frac{\hat{\theta}(X_1, \ldots, X_n) - \mathrm{E}_\theta\, \hat{\theta}(X_1, \ldots, X_n)}{\sqrt{\mathrm{Var}_{\,\theta}\, \hat{\theta}(X_1, \ldots, X_n)}} \xrightarrow[n\to\infty]{d} Y \sim N(0, 1)\,,$$

dann ist $\hat{\theta}(X_1, \ldots, X_n)$ *asymptotisch normalverteilt*.

Definition 8.3.7 (Bester erwartungstreuer Schätzer) Der Punktschätzer $\hat{\theta}(X_1, \ldots, X_n)$ für θ ist der *beste erwartungstreue Schätzer*, falls

$$\mathrm{E}_\theta\, \hat{\theta}^2(X_1, \ldots, X_n) < \infty\,, \quad \theta \in \Theta\,, \qquad \mathrm{E}_\theta\, \hat{\theta}(X_1, \ldots, X_n) = \theta\,, \quad \theta \in \Theta\,,$$

und $\hat{\theta}$ die minimale Varianz in der Klasse aller erwartungstreuen Schätzer für θ besitzt. Das heißt, dass für einen beliebigen erwartungstreuen Schätzer $\tilde{\theta}(X_1, \ldots, X_n)$ mit

$$\mathrm{E}_\theta\, \tilde{\theta}^2(X_1, \ldots, X_n) < \infty \quad \text{gilt} \quad \mathrm{Var}_{\,\theta}\, \hat{\theta} \le \mathrm{Var}_{\,\theta}\, \tilde{\theta}\,, \qquad \theta \in \Theta\,.$$

8.3.2 Schätzer des Erwartungswertes und empirische Momente

Sei $X \stackrel{d}{=} X_i\,, \quad i = 1, \ldots, n$ ein statistisches Merkmal. Sei weiter $\mathrm{E}|X_i|^k < \infty$ für ein $k \in \mathbb{N}$, $m = 1$ und der zu schätzende Parameter $\theta = \mu_k = \mathrm{E}X_i^k$. Insbesondere gilt im Fall $k = 1$, dass $\theta = \mu_1 = \mu$ der Erwartungswert ist.

Definition 8.3.8 Das *k-te empirische Moment* von X wird als

$$\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

definiert. Unter dieser Definition gilt, dass $\hat{\mu}_1 = \bar{X}_n$, also das erste empirische Moment gleich dem Stichprobenmittel ist.

Satz 8.3.1 (Eigenschaften der empirischen Momente) Unter obigen Voraussetzungen gelten folgende Eigenschaften:

1. $\hat{\mu}_k$ ist erwartungstreu für μ_k (insbesondere $\bar{X}_n$).
2. $\hat{\mu}_k$ ist stark konsistent.
3. Falls $\mathrm{E}_\theta |X|^{2k} < \infty\,, \quad \forall \theta \in \Theta$, dann ist $\hat{\mu}_k$ asymptotisch normalverteilt.
4. Es gilt Var $\bar{X}_n = \frac{\sigma^2}{n}$, wobei $\sigma^2 = \mathrm{Var}_{\,\theta}\, X$. Falls $X_i \sim N(\mu, \sigma^2)\,, \quad i = 1, \ldots, n$ (eine normalverteilte Stichprobe), dann gilt:

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)\,.$$

8.3.3 Schätzer der Varianz

Seien $X_i\,, \quad i = 1, \ldots, n$ unabhängig identisch verteilt, $X_i \stackrel{d}{=} X\,, \quad \mathrm{E}_\theta\, X^2 < \infty \quad \forall \theta \in \Theta\,, \quad \theta = (\theta_1, \ldots, \theta_m)^T\,,\ \theta_i = \sigma^2 = \mathrm{Var}_{\,\theta}\, X$ für ein $i \in \{1, \ldots, m\}$. Die *Stichprobenvarianz*

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

ist dann ein Schätzer für σ^2. Falls der Erwartungswert $\mu = \mathrm{E}_\theta X$ der Stichprobenvariablen explizit benannt ist, so kann ein Schätzer für σ^2 auch als

$$\tilde{S}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$$

definiert werden.
Wir werden nun die Eigenschaften von S_n^2 und $\tilde{S}_n^2$ untersuchen und sie miteinander vergleichen.

Satz 8.3.2 1. Die Stichprobenvarianz S_n^2 ist erwartungstreu für σ^2:

$$\mathrm{E}_\theta\, S_n^2 = \sigma^2\,, \qquad \theta \in \Theta\,.$$

2. Wenn $\mathrm{E}_\theta\, X^4 < \infty$, dann gilt

$$\mathrm{Var}_{\,\theta}\, S_n^2 = \frac{1}{n}\left(\mu_4' - \frac{n-3}{n-1}\sigma^4\right),$$

wobei $\mu_4' = \mathrm{E}_\theta\,(X-\mu)^4$.

Satz 8.3.3 1. Der Schätzer $\tilde{S}_n^2$ für σ^2 ist erwartungstreu.

2. Es gilt $\mathrm{Var}_{\,\theta}\, \tilde{S}_n^2 = 1/n(\mu_4' - \sigma^4)$.

Folgerung 8.3.1 Der Schätzer $\tilde{S}_n^2$ für σ^2 ist besser als S_n^2, weil beide erwartungstreu sind und

$$\mathrm{Var}_{\,\theta}\, \tilde{S}_n^2 = \frac{\mu_4' - \sigma^4}{n} < \frac{\mu_4' - \frac{n-3}{n-1}\sigma^4}{n} = \mathrm{Var}_{\,\theta}\, S_n^2\,.$$

Diese Eigenschaft von $\tilde{S}_n^2$ im Vergleich zu S_n^2 ist intuitiv klar, da man in $\tilde{S}_n^2$ mehr Informationen über die Verteilung der Stichprobenvariablen X_i (nämlich den bekannten Erwartungswert μ) reingesteckt hat.

Satz 8.3.4 Die Schätzer S_n^2 bzw. $\tilde{S}_n^2$ sind stark konsistent und asymptotisch normalverteilt:

$$S_n^2 \xrightarrow[n\to\infty]{\text{f.s.}} \sigma^2\,, \qquad \sqrt{n}\frac{S_n^2 - \sigma^2}{\sqrt{\mu_4' - \sigma^4}} \xrightarrow[n\to\infty]{d} Y \sim N(0,1)\,,$$

$$\tilde{S}_n^2 \xrightarrow[n\to\infty]{\text{f.s.}} \sigma^2\,, \qquad \sqrt{n}\frac{\tilde{S}_n^2 - \sigma^2}{\sqrt{\mu_4' - \sigma^4}} \xrightarrow[n\to\infty]{d} Y \sim N(0,1)\,,$$

falls $\mu_4' < \infty$.

Folgerung 8.3.2 Es gilt

1.
$$\sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \xrightarrow[n\to\infty]{d} Y \sim N(0,1)$$

und somit

2.
$$P\left(\mu \in \left[\bar{X}_n - \frac{z_{1-\alpha/2}S_n}{\sqrt{n}}\,,\, \bar{X}_n + \frac{z_{1-\alpha/2}S_n}{\sqrt{n}}\right]\right) \xrightarrow[n\to\infty]{} 1-\alpha \qquad (8.3)$$

für ein $\alpha \in (0,1)$, wobei z_α das α-Quantil der $N(0,1)$-Verteilung ist.

Bemerkung 8.3.2 Das Intervall in (8.3) wird *asymptotisches Konfidenz- oder Vertrauensintervall* für den Parameter μ genannt. Falls α klein ist (z.B. $\alpha = 0,05$), so liegt μ mit einer asymptotisch großen Wahrscheinlichkeit $1-\alpha$ im vorgegebenen Intervall. Diese Art der Schätzung von μ stellt eine Alternative zu den Punktschätzern dar und wird ausführlich in Kapitel 4 behandelt.

Betrachten wir weiterhin den wichtigen Spezialfall der normalverteilten Stichprobenvariablen $X_i\,, \quad i = 1, \ldots, n$, also $X \sim N(\mu, \sigma^2)$.

Satz 8.3.5 Falls $X_1, \ldots, X_n$ normalverteilt sind mit Parametern μ und σ^2, dann gilt

1. $$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}\,,$$
2. $$\frac{n\tilde{S}_n^2}{\sigma^2} \sim \chi^2_n\,.$$

Lemma 8.3.2 Falls $X \sim N(\mu, \sigma^2)$, $X_1, \ldots, X_n$ unabhängige identisch verteilte Zufallsvariablen, $X_i \stackrel{d}{=} X$, dann sind $\bar{X}_n$ und S_n^2 unabhängig.

Dieses Lemma wird unter Anderem gebraucht, um folgendes Ergebnis zu beweisen:

Satz 8.3.6 Unter den Voraussetzungen von Lemma 8.3.2 gilt

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \sim t_{n-1}\,.$$

Bemerkung 8.3.3 Es sei $(X_1, \ldots, X_n)$ eine normalverteilte Stichprobe, $X_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$. Mit Hilfe des Satzes 8.3.6 kann folgendes Konfidenzintervall für den Erwartungswert μ konstruiert werden:

$$P\left(\mu \in \left[\bar{X}_n - \frac{t_{n-1,1-\alpha/2}}{\sqrt{n}} S_n\,,\; \bar{X}_n + \frac{t_{n-1,1-\alpha/2}}{\sqrt{n}} S_n\right]\right) = 1-\alpha$$

für $\alpha \in (0,1)$, denn

$$\begin{aligned} P\left(\sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \in \left[\underbrace{t_{n-1,\alpha/2}}_{=-t_{n-1,1-\alpha/2} \text{ wg. Sym. } t\text{-Vert.}}\,,\; t_{n-1,1-\alpha/2}\right]\right) &= F_{t_{n-1}}(t_{n-1,1-\alpha/2}) - F_{t_{n-1}}(t_{n-1,\alpha/2}) \\ &= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1-\alpha\,, \end{aligned} \tag{8.4}$$

wobei $t_{n-1,\alpha}$ das α-Quantil der t_{n-1}-Verteilung darstellt. Der Rest folgt aus (8.4) durch das Auflösen bzgl. μ.

8.3.4 Eigenschaften der Ordnungsstatistiken

In Abschnitt 7.6.2 haben wir bereits die Ordnungsstatistiken $x_{(1)}, \ldots, x_{(n)}$ einer konkreten Stichprobe $(x_1, \ldots, x_n)$ betrachtet. Wenn wir nun auf der Modellebene arbeiten, also eine Zufallsstichprobe $(X_1, \ldots, X_n)$ von unabhängigen identisch verteilten Zufallsvariablen X_i mit Verteilungsfunktion $F(x)$ haben, welche Eigenschaften haben dann ihre Ordnungsstatistiken

$$X_{(1)}, \ldots, X_{(n)}\,?$$

Satz 8.3.7

1. Die Verteilungsfunktion der Ordnungsstatistik $X_{(i)}\,, \quad i = 1, \ldots, n$ ist gegeben durch

$$F_{X_{(i)}}(x) = \sum_{k=i}^{n} \binom{n}{k} F^k(x)(1 - F(x))^{n-k}\,, \qquad x \in \mathbb{R}\,. \tag{8.5}$$

2. Falls X_i absolut stetig verteilt sind mit Dichte f, die stückweise stetig ist, dann ist auch $X_{(i)}$, $i = 1, \ldots, n$ absolut stetig verteilt mit der Dichte

$$f_{X_{(i)}}(x) = \frac{n!}{(i-1)!(n-i)!} f(x) F^{i-1}(x)(1 - F(x))^{n-i}\,, \quad x \in \mathbb{R}\,.$$

Beweis

Führen wir die Zufallsvariable

$$Y = \#\{i : X_i \le x\} = \sum_{i=1}^{n} I(X_i \le x)\,, \quad x \in \mathbb{R}$$

ein. Da $X_1, \ldots, X_n$ unabhängig identisch verteilt mit Verteilungsfunktion F sind, gilt $Y \sim Bin(n, F(x))$. Weiterhin gilt

$$F_{X_{(i)}}(x) = P(X_{(i)} \le x) = P(Y \ge i) = \sum_{k=i}^{n} \binom{n}{k} F^k(x)(1-F(x))^{n-k}\,, \quad x \in \mathbb{R}\,.$$

□

Bemerkung 8.3.4 Für $i = 1$ und $i = n$ sieht die Formel (8.5) besonders einfach aus:

$$F_{X_{(1)}}(x) = 1 - (1 - F(x))^n\,, \qquad x \in \mathbb{R}$$
$$F_{X_{(n)}}(x) = F^n(x)\,, \qquad x \in \mathbb{R}\,.$$

Übungsaufgabe 8.3.1 Zeigen Sie für $X_1, \ldots, X_n$ unabhängig identisch verteilt, $X_i \sim U[0, \theta]$, $\theta > 0$, $i = 1, \ldots, n$, dass

1. die Dichte von $X_{(i)}$ gleich

$$f_{X_{(i)}}(x) = \begin{cases} \frac{n!}{(i-1)!(n-i)!}\theta^{-n}x^{i-1}(\theta - x)^{n-i}\,, & x \in (0,\theta) \\ 0\,, & \text{sonst} \end{cases}$$

und

2.

$$\mathrm{E}X_{(i)}^k = \frac{\theta^k n!(i+k-1)!}{(n+k)!(i-1)!}\,, \qquad k \in \mathbb{N}\,, \quad i = 1,\ldots,n$$

sind. Insbesondere gilt $\mathrm{E}X_{(i)} = \frac{i}{n+1}\theta$ und Var $X_{(i)} = \frac{i(n-i+1)\theta^2}{(n+1)^2(n+2)}$.

8.3.5 Empirische Verteilungsfunktion

Im Folgenden betrachten wir die statistischen Eigenschaften der in Abschnitt 7.5.2 eingeführten empirischen Verteilungsfunktion $\hat{F}_n(x)$ einer Zufallsstichprobe $(X_1,\ldots,X_n)$, wobei $X_i \stackrel{d}{=} X$ unabhängige identisch verteilte Zufallsvariablen mit Verteilungsfunktion $F(\cdot)$ sind.

Satz 8.3.8 Es gilt

1. $n\hat{F}_n(x) \sim Bin(n, F(x))\,, \quad x \in \mathbb{R}.$
2. $\hat{F}_n(x)$ ist ein erwartungstreuer Schätzer für $F(x)$, $x \in \mathbb{R}$ mit

$$\text{Var } \hat{F}_n(x) = \frac{F(x)(1-F(x))}{n}\,.$$

3. $\hat{F}_n(x)$ ist stark konsistent.
4. $\hat{F}_n(x)$ ist asymptotisch normalverteilt:

$$\sqrt{n}\frac{\hat{F}_n(x) - F(x)}{\sqrt{F(x)(1-F(x))}} \xrightarrow{d} Y \sim N(0,1)\,, \qquad \forall x:\, F(x) \in (0,1)\,.$$

In Satz 8.3.8, 3) wird behauptet, dass

$$\hat{F}_n(x) \xrightarrow[n\to\infty]{\text{f.s.}} F(x)\,, \qquad \forall x \in \mathbb{R}\,.$$

Der nachfolgende Satz von Gliwenko-Cantelli behauptet, dass diese Konvergenz gleichmäßig in $x \in \mathbb{R}$ stattfindet. Um diesen Satz formulieren zu können, betrachten wir den *gleichmäßigen Abstand* zwischen $\hat{F}_n$ und F

$$D_n = \sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)|\,.$$

Dieser Abstand ist eine Zufallsvariable, die auch *Kolmogorow-Abstand* genannt wird. Er gibt den maximalen Fehler an, den man bei der Schätzung von $F(x)$ durch $\hat{F}_n(x)$ macht.

Übungsaufgabe 8.3.2 Zeigen Sie, dass

$$D_n = \max_{i\in\{1,\dots,n\}} \max\left\{F\left(X_{(i)}-0\right) - \frac{i-1}{n}\,,\, \frac{i}{n} - F\left(X_{(i)}\right)\right\}\,. \tag{8.6}$$

Beachten Sie dabei die Tatsache, dass $\hat{F}_n(x)$ eine Treppenfunktion mit Sprungstellen $X_{(i)}$, $i = 1,\dots,n$ ist.

Satz 8.3.9 (Gliwenko-Cantelli) Es gilt $D_n \xrightarrow[n\to\infty]{\text{f.s.}} 0$.

Satz 8.3.10 Für jede stetige Verteilungsfunktion F gilt

$$D_n \overset{d}{=} \sup_{y\in[0,1]} \left|\hat{G}_n(y) - y\right|\,,\ \text{wobei } \hat{G}_n(y) = \frac{1}{n}\sum_{i=1}^{n} I\left(Y_i \le y\right)\,,\quad y\in\mathbb{R}$$

die empirische Verteilungsfunktion der Zufallsstichprobe $(Y_1,\dots,Y_n)$ mit unabhängigen identisch verteilten Zufallsvariablen $Y_i \sim U[0,1]\,,\quad i = 1,\dots,n$ ist.

Folgerung 8.3.3 Falls F eine stetige Verteilungsfunktion ist, dann gilt

$$D_n \overset{d}{=} \max_{i=1,\dots,n} \max\left\{Y_{(i)} - \frac{i-1}{n}\,,\, \frac{i}{n} - Y_{(i)}\right\}\,,$$

wobei $Y_{(1)},\dots,Y_{(n)}$ die Ordnungsstatistiken der auf $[0,1]$ gleichverteilten Stichprobenvariablen $Y_1,\dots,Y_n$ sind.

Beweis Benutze dazu die Darstellung (8.6), den Satz 8.3.10 sowie die Tatsache, dass

$$F(x) = x\,,\quad x\in[0,1]$$

für die Verteilungsfunktion der $U[0,1]$-Verteilung ist. □

Folgende Ergebnisse werden ohne Beweis angegeben:

Bemerkung 8.3.5 Für die Zwecke des statistischen Testens (vgl. den Anpassungstest von Kolmogorow-Smirnow, Bemerkung 8.3.6, 2)) ist es notwendig, die Quantile der Verteilung von D_n zu nennen. Auf Grund der Komplexität der Verteilung von D_n ist es jedoch unmöglich, sie explizit anzugeben. Mit Hilfe des Satzes 8.3.10 ist es möglich, diese Quantile durch Monte-Carlo-Simulationen numerisch zu berechnen. Dazu simuliert man mehrere Stichproben $(Y_1,\dots,Y_n)$ von $U[0,1]$-verteilten Pseudozufallszahlen, bildet $\hat{G}_n(x)$ und berechnet D_n nach Folgerung 8.3.3.

Satz 8.3.11 (Kolmogorow) Falls die Verteilungsfunktion F der unabhängigen und identisch verteilten Stichprobenvariablen X_i, $i = 1,\dots,n$ stetig ist, dann gilt

$$\sqrt{n}D_n \xrightarrow[n\to\infty]{d} Y\,,$$

wobei Y eine Zufallsvariable mit der Verteilungsfunktion

$$K(x) = \begin{cases} \sum_{k=-\infty}^{\infty}(-1)^k e^{-2k^2x^2} = 1 + 2\sum_{k=1}^{\infty}(-1)^k e^{-2k^2x^2}\,, & x > 0\,, \\ 0\,, & \text{sonst} \end{cases}$$

(Kolmogorow-Verteilung) ist.

Bemerkung 8.3.6 1. Aus Satz 8.3.11 folgt

$$P(\sqrt{n}D_n \le x) \underset{n\to\infty}{\approx} K(x)\,, \quad x \in \mathbb{R}\,.$$

Die daraus resultierende Näherungsformel

$$P(D_n \le x) \approx K(x\sqrt{n})$$

ist ab $n > 40$ praktisch brauchbar.

2. *Kolmogorow-Smirnow-Anpassungstest:* Mit Hilfe der Aussage des Satzes 8.3.11 ist es möglich, folgenden *asymptotischen Anpassungstest von Komogorow-Smirnow* zu entwickeln. Es wird die Haupthypothese $H_0 : F = F_0$ (die unbekannte Verteilungsfunktion der Stichprobenvariablen $X_1, \ldots, X_n$ ist gleich F_0) gegen die Alternative $H_1 : F \neq F_0$ getestet. Dabei wird H_0 verworfen, falls

$$\sqrt{n}D_n \notin [k_{\alpha/2}\,,\, k_{1-\alpha/2}]$$

ist, wobei

$$D_n = \sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F_0(x)|$$

und k_α das α-Quantil der Kolmogorow-Verteilung ist. Somit ist die Wahrscheinlichkeit, die richtige Hypothese H_0 zu verwerfen (Wahrscheinlichkeit des *Fehlers 1. Art*) asymptotisch gleich

$$P\left(\sqrt{n}D_n \notin [k_{\alpha/2}\,,\, k_{1-\alpha/2}]\,|\,H_0\right) \underset{n\to\infty}{\longrightarrow} 1 - K(k_{1-\alpha/2}) + K(k_{\alpha/2}) = 1 - (1 - \alpha/2) + \alpha/2 = \alpha\,.$$

In der Praxis wird α klein gewählt, z.B. $\alpha \approx 0,05$. Somit ist im Fall, dass H_0 stimmt, die Wahrscheinlichkeit einer Fehlentscheidung in Folge des Testens klein.

Dieser Test ist nur ein Beispiel dessen, wie der Satz von Kolmogorow in der statistischen Testtheorie verwendet wird. Die allgemeine Philosophie des Testens wird in Stochastik III erläutert.

Mit Hilfe von $\hat{F}_n$ lassen sich sehr viele Schätzer durch die sogenannte *Plug-in-Methode* konstruieren. Dies werden wir jetzt näher erläutern: Sei $M = \{\text{Menge aller Verteilungsfunktionen}\}$.

Literaturverzeichnis

[1] S. Asmussen and P. W. Glynn. *Stochastic simulation: algorithms and analysis*, volume 57 of *Stochastic Modelling and Applied Probability*. Springer, New York, 2007.

[2] H. Dehling, B. Haupt. *Einführung in die Wahrscheinlichkeitstheorie und Statistik*. Springer, Berlin, 2003.

[3] H. Bauer. *Wahrscheinlichkeitstheorie*. de Gruyter, Berlin, 1991.

[4] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. C. Van der Meulen. Nonparametric entropy estimation: an overview. *Int. J. math. Stat. Sci.*, 6(1):17–39, 1997.

[5] A. A. Borovkov. *Wahrscheinlichkeitstheorie: eine Einführung*. Birkhäuser, Basel, 1976.

[6] J. A. Costa and A. O. Hero, III. Determining intrinsic dimension and entropy of high-dimensional shape spaces. In *Statistics and analysis of shapes*, Model. Simul. Sci. Eng. Technol., pages 231–252. Birkhäuser Boston, Boston, MA, 2006.

[7] C. D. Daykin, T. Pentikäinen, and M. Pesonen. *Practical risk theory for actuaries*, volume 53 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, Ltd., London, 1994.

[8] W. Feller. *An introduction to probability theory and its applications. Vol I/II*. J. Wiley & Sons, New York, 1970/71.

[9] H. O. Georgii. *Stochastik*. de Gruyter, Berlin, 2002.

[10] B. V. Gnedenko. *Einführung in die Wahrscheinlichkeitstheorie*. Akademie, Berlin, 1991.

[11] C. Graham and D. Talay. *Stochastic simulation and Monte Carlo methods*, volume 68 of *Stochastic Modelling and Applied Probability*. Springer, Heidelberg, 2013. Mathematical foundations of stochastic simulation.

[12] C. Hesse. *Angewandte Wahrscheinlichkeitstheorie*. Vieweg, Braunschweig, 2003.

[13] M. L. Huber. *Perfect simulation*, volume 148 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL, 2016.

[14] A. F. Karr. *Probability*. Springer, New York, 1993.

[15] L. F. Kozachenko and N. N. Leonenko. A statistical estimate for the entropy of a random vector. *Problemy Peredachi Informatsii*, 23(2):9–16, 1987.

[16] U. Krengel. *Einführung in die Wahrscheinlichkeitstheorie*. Vieweg, Braunschweig, 2002.

[17] D. P. Kroese, T. Taimre, and Z. I. Botev. *Handbook of Monte Carlo methods*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, 2011.

[18] N. Leonenko, L. Pronzato, and V. Savani. A class of Rényi information estimators for multidimensional densities. *Ann. Statist.*, 36(5):2153–2182, 2008.

[19] N. Leonenko, L. Pronzato, and V. Savani. Estimation of entropies and divergences via nearest neighbors. *Tatra Mt. Math. Publ.*, 39:265–273, 2008.

[20] N. Metropolis and S. Ulam. The Monte Carlo method. *J. Amer. Statist. Assoc.*, 44:335–341, 1949.

[21] B. L. Nelson. *Foundations and methods of stochastic simulation*, volume 187 of *International Series in Operations Research & Management Science*. Springer, New York, 2013. A first course.

[22] M. Nilsson and W. B. Kleijn. On the estimation of differential entropy from data located in embedded manifolds. *IEEE Trans. Inform. Theroy*, 57(7):2330–2341, 2007.

[23] J. Jacod, P. Protter. *Probability essentials*. Springer, Berlin, 2003.

[24] H. H. Panjer. Recursive evaluation of a family of compound distributions. *Astin Bull.*, 12(1):22–26, 1981.

[25] M. D. Penrose and J. E. Yukich. Limit theory for point processes in manifolds. *Ann. Appl. Probab.*, 23(6):2161–2211, 2013.

[26] A. Rényi. A few fundamental problems of information theory. *Magyar Tud. Akad. Mat. Fiz. Oszt. Közl.*, 10:251–282, 1960.

[27] A. Rényi. On measures of entropy and information. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I*, pages 547–561. Univ. California Press, Berkeley, Calif., 1961.

[28] T. Rolski, H. Schmidli, V. Schmidt, and J. Teugels. *Stochastic processes for insurance and finance*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 1999.

[29] S. M. Ross. *Simulation*. Elsevier/Academic Press, Amsterdam, 2013. Fifth edition.

[30] R. Y. Rubinstein and D. P. Kroese. *Simulation and the Monte Carlo method*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, 2017. Third edition.

[31] L. Sachs. *Angewandte Statistik*. Springer, 2004.

[32] A. N. Shiryaev. *Probability*. Springer, New York, 1996.

[33] K.-S. Song. Rényi information, loglikelihood and an intrinsic distribution measure. *J. Statist. Plann. Inference*, 93(1-2):51–69, 2001.

[34] J. M. Stoyanov. *Counterexamples in probability*. Wiley & Sons, 1987.

[35] B. Sundt and W. S. Jewell. Further results on recursive evaluation of compound distributions. *Astin Bull.*, 12(1):27–39, 1981.

[36] N. T. Thomopoulos. *Essentials of Monte Carlo simulation*. Springer, New York, 2013.

[37] H. Tijms. *Understanding probability. Chance rules in everyday life.* Cambridge University Press, 2004.

[38] P. Gänßler, W. Stute. *Wahrscheinlichkeitstheorie*. Springer, Berlin, 1977.